Double Reading in Breast Cancer Screening: Cohort Evaluation in the CO-OPS Trial[1]

Sian Taylor-Phillips, PhD
David Jenkinson, PhD
Chris Stinton, PhD
Matthew G. Wallis, FRCR
Janet Dunn, PhD
Aileen Clarke, MD

Purpose: To investigate the effect of double readings by a second radiologist on recall rates, cancer detection, and characteristics of cancers detected in the National Health Service Breast Screening Program in England.

Materials and Methods: In this retrospective analysis, 805 206 women were evaluated through screening and diagnostic test results by extracting 1 year of routine data from 33 English breast screening centers. Centers used double reading of digital mammograms, with arbitration if there were discrepant interpretations. Information on reader decisions, with results of follow-up tests, were used to explore the effect of the second reader. The statistical tests used were the test for equality of proportions, the χ^2 test for independence, and the *t* test.

Results: The first reader recalled 4.76% of women (38 295 of 805 206 women; 95% confidence interval [CI]: 4.71%, 4.80%). Two readers recalled 6.19% of women in total (49 857 of 805 206 women; 95% CI: 6.14%, 6.24%), but arbitration of discordant readings reduced the recall rate to 4.08% (32 863 of 805 206 women; 95% CI: 4.04%, 4.12%; $P < .001$). A total of 7055 cancers were detected, of which 627 (8.89%; 95% CI: 8.22%, 9.55%; $P < .001$) were detected by the second reader only. These additional cancers were more likely to be ductal carcinoma in situ (30.5% [183 of 600] vs 22.0% [1344 of 6114]; $P < .001$), and additional invasive cancers were smaller (mean size, 14.2 vs 16.7 mm; $P < .001$), had fewer involved nodes, and were likely to be lower grade.

Conclusion: Double reading with arbitration reduces recall and increases cancer detection compared with single reading. Cancers detected only by the second reader were smaller, of lower grade, and had less nodal involvement.

[1] From the Warwick Medical School, University of Warwick, Gibbett Hill Rd, Coventry CV4 7AL, England (S.T., D.J., C.S., J.D., A.C.); and Cambridge Breast Unit, Cambridge University Hospitals National Health Service Foundation Trust, and National Institute for Health Research Cambridge Biomedical Research Centre, Cambridge, England (M.G.W.). Received May 14, 2017; revision requested June 23; final revision received January 5, 2018; accepted January 5, 2018. **Address correspondence to** S.T.P. (e-mail: *S.Taylor-Phillips@warwick.ac.uk).*

This article presents independent research funded by the National Institute for Health Research (NIHR). S.T. supported by the NIHR Career Development Fellowship Programme (CDF-2016-09-018, wmclahrc-2014-1). S.T., D.J., C.S., and A.C. supported by the NIHR Collaboration for Leadership in Applied Health Research and Care West Midlands. The sponsor is the University of Warwick. The views expressed are those of the authors and not necessarily those of the National Health Service, the NIHR, or the Department of Health and Social Care.

Breast cancer is a leading cause of cancer in women (1), and many countries have implemented screening programs. Despite concerns about the balance of benefits and harms of these programs, results of randomized controlled trials indicate that screening reduces mortality from breast cancer (2,3).

In many European countries, the interpretation of mammograms is performed by two readers. Recall occurs if *(a)* either reader suggests it, *(b)* through consensus, or *(c)* after arbitration by a third (or more) additional readers (4,5). In the United States, mammograms are typically interpreted by a single reader accompanied by computer-aided detection (6). There is debate about the benefits and costs of single versus double reader programs. Some film mammography studies indicate that double reading increases the number of cancers detected (7–12) but results in the recall of more women (9,11–14) and requires more resources (15). It might increase detection of small (<15 mm) cancers (16) and identify a higher ratio of ductal carcinoma in situ (DCIS) to invasive cancers (15), which is potentially undesirable because of the association between DCIS and overdiagnosis (17,18). Other studies report no differences in the size or stage of cancers between single and double reader programs (9,19).

Digital mammography has replaced film mammography in routine clinical practice (20,21). Yet despite the widespread use of digital mammography and double reading, there is little published data on their combined effects. Three small studies, and a meta-analysis of these studies, found no statistically significant difference in cancer detection rates between single and double reader strategies (22–25). Extra cancers identified by second readers were more likely to be DCIS than invasive carcinomas (23). Posso and colleagues (23,24) have tentatively suggested that single reader screening could reduce costs in breast cancer programs without decreasing cancer detection rates. However, results may be because of small sample sizes, as the second reader detected an extra 10% (*n* = 24) cancers, but this was not statistically significant (24).

The key limitations of the evidence base are that most data come from film mammography studies (which does not reflect modern breast screening) and that studies using digital mammography have had relatively small samples (maximum = 57 157) and detected few cancers (limiting their power to detect differences). Our purpose was to examine the impact of double reading on recall and cancer detection rates and the characteristics of the additional cancers identified by double reading in the National Health Service Breast Screening Program in England.

Implications for Patient Care

- Use of a second radiology reader in breast cancer screening can increase the number of cancers detected; however, the clinical importance of these cancers requires careful consideration as some may be overdiagnosed.
- The additional cancers detected by only the second reader were more likely to be ductal carcinoma in situ and lower-grade tumors.
- Double reading with arbitration of discordant examinations resulted in recall of fewer women for further tests than would have been recalled if only the first reader decision was used.
- Use of effective arbitration of discordant examinations can reduce the number of women recalled for further tests.

Materials and Methods

Study Design and Participants

This is a population-based cohort study nested within the Changing Case Order to Optimize Patterns of Performance in Screening (CO-OPS) Trial, which included 1 194 147 women between 47 and 73 years of age at 46 screening centers, all between December 20, 2012, and November 3, 2014 (ISRCTN46603370, ethical approvals: Coventry and Warwickshire National Health Service [NHS] Research Ethics Committee, June 27, 2012, WM/0182) (26). Each center participated for a year, and every woman screened as part of the United Kingdom National Health Service population breast screening program was included in this study. Women with symptoms at presentation and women who were tested because of familial or other risk factors were excluded. In this cohort study, we analyzed data from 33 centers: 13 centers were excluded because they used arbitration after both readers agreed to recall. A total of 805 665 women were included in the analysis, all of whom have previously been reported in an analysis of radiologist performance with time on task (25) but none of whom have previously been reported in a comparison of single and double reading.

Procedures

In the United Kingdom, women between the ages of 50 and 70 years are invited to mammographic screening every 3 years, with a trial of age extension from 47 to 73 years. Two views of each breast are obtained, mediolateral oblique and craniocaudal. Mammograms are reviewed by two readers from the same

https://doi.org/10.1148/radiol.2018171010

Content code: BR

Radiology 2018; 287:749–757

Abbreviations:
DCIS = ductal carcinoma in situ
NHS = National Health Service

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

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

See also the editorial by Rosenberg and Seidenwurm.

breast screening center using digital mammography without computer-aided detection. They are instructed to read batches of women's mammograms independently but can view the other readers' decisions in patient records. They are aware of whether they are the first or second reader in the workflow processes. Twelve of 33 centers used workflow systems designed to blind the second reader to the decision of the first reader. Disagreements between readers were resolved either through a single third reader (n = 11 centers) or by group consensus (n = 22 centers). Arbitration was performed by qualified readers from the same screening center. All readers were accredited by the National Health Service Breast Screening Program; readers undergo formal training, read a minimum of 5000 women's mammograms per year, participate in assessment clinics, audit their own performance, and maintain continuing professional development (4). Each service is expected to perform within set parameters, including cancer detection and recall rates (4). Readers take 35 seconds on average to examine each woman's digital mammograms in the NHS Breast Screening Program (27). Women recalled after screening are offered further tests at assessment, according to national guidelines (28).

Outcomes

The main outcomes were recall and cancer detection rates. Cancer was defined as histologically confirmed invasive cancer or DCIS. Absence of cancer was confirmed either through arbitration by expert readers or follow-up tests including ultrasonography (US), magnetic resonance (MR) imaging, and biopsy. Where 3-year follow-up data were available (for women attending 10 of the first centers to complete the trial), interval cancer rates between screening rounds and cancer detection rates at the following screening were measured as an alternative reference standard to determine the absence of cancer. Secondary outcomes were characteristics of cancers detected, specifically the proportion that included any invasive cancer (rather than DCIS only), the grade, the number of involved nodes, the pathologic size for women with invasive cancer, and the grade for women with DCIS only.

Data Collection

Data were extracted from the National Breast Screening Service electronic database. We extracted the decisions of the first and second readers (and arbitration, where used) for whether the patient should be recalled for further tests, which are recorded automatically at the point of making the decision. The decision of arbitration was final, and to confirm this we checked against records scheduling the follow-up appointments. For all follow-up appointments, we extracted whether the woman had a biopsy, the biopsy result (pathologic finding), and the result of other follow-up tests used (additional mammography, clinical breast examination, US, and/or MR imaging). We extracted pathologic results after any subsequent surgery. This was used to confirm biopsy results and to report grade, size, and number of involved nodes. We extracted interval cancer rates between screening rounds and cancer detection rates at the following screening round 3 years later for women attending 10 of the first centers to complete the trial, as in these centers sufficient time has elapsed to extract these data. This was used to investigate whether women with discordant readings who were not recalled after arbitration were at increased risk for later cancer detection (which may be an indication of errors in arbitration and potential underestimation of the extra cancers detected by the second reader).

Statistical Analysis

The analysis compared the recall rate from three screening approaches. The first (double reading plus arbitration) was what was used in clinical practice. Two readers independently examined each case and indicated whether they think the woman should be recalled for further tests. If they disagreed, then expert arbitration was used to make the final decision. The second approach (single reader) derives results from whether the first reader alone judged that the woman should be recalled. The third (recall if either reader suggests) counts every woman recalled by either reader as recalled.

The number of cancers detected with double reading plus arbitration was compared with the number detected by reader 1 alone. The characteristics of the extra cancers detected by the second reader alone (missed by the first reader) were compared with those of cancers detected by reader 1. The number of involved nodes was grouped into none, one to two, and three or more, as these categories relate to prognosis. The statistical tests used were the test for equality of proportions, the χ^2 test for independence, and the t test. We performed a sensitivity analysis assuming all missing data were extreme cases (invasive disease not present, lowest grade, without nodal involvement, or vice versa). The analysis was performed by using R statistical software, version 3.4.1, in RStudio, version 1.0.153 (29). For women at 10 of the first centers to complete the trial, we report the 3-year interval cancer rate and cancer detection rate at their subsequent screening examination 3 years later. These results were divided into three groups: Women who were recalled by the first reader but not by the second reader and arbitration at the current screening, women who were recalled by the second reader but not by the first reader and arbitration at the current screening, and all other women who were not recalled at the current screening (recalled by neither reader). Comparisons between these groups were made by using the test for equality of proportions. If women who had a discordant reading but who were not recalled at the current round had a higher cancer detection rate in the subsequent round, this may indicate that arbitration was incorrect and cancers were missed at the current round. However, it may also be caused by discordant cases having other risk factors for developing cancers between screening rounds, such as increased breast density. As a sensitivity analysis, the number of additional cancers detected by the second reader was recalculated assuming that the difference between cancer detection rate in discordant and nondiscordant readings at the subsequent round was due entirely to cancers missed by arbitration at the current round, and that the differences at the 10 centers would not differ from those across the whole data set.

Results

The flow of women through the study is detailed in Figure 1. Of the 805 665 women screened, 805 206 had complete records of first and second reader screening decisions. A total of 459 women (0.1%) were excluded from further analysis because 44 were examined by a single reader only and recalled for further tests, and 425 were examined by a single reader only and not recalled for further tests. All women had complete records for whether they were recalled for further tests and for whether the results of those further tests showed any type of cancer (DCIS or invasive). The median age of the women included was 59 years (interquartile range, 53–65 years), and 169 753 women (21.1%) were attending their first ever screening appointment.

A total of 7055 cancers were detected. Details of missing data are provided in the Table. Excluding missing data, invasive disease was present in 77.3% (5190 of 6717) of cancers, of which 20.4% (1048 of 5147) were grade 3, 54.3% (2797 of 5147) were grade 2, and 25.3% (1302 of 5147) were grade 1, and the mean pathologic size was 16.5 mm ± 12.1 (standard deviation) (n = 5180). DCIS alone was present in 22.7% (1527 of 6717) of cancers, of which 63.1% (830 of 1316) were high grade, 27.1% (356 of 1316) were intermediate grade, and 9.9% (130 of 1316) were low grade. For cancers with invasive disease present, 76.7% (3909 of 5097) were axillary node negative, 17.3% (884 of 5097) had one or two nodes involved, and only 6.0% (304 of 5097) had three or more nodes involved.

Comparison between Different Single and Double Readings

The recall rate was 4.08% (32 863 of 805 206; 95% confidence interval [CI]: 4.04%, 4.12%) after arbitration. In comparison, if there were no arbitration and women were recalled if either reader suggested it, the recall rate would have been 6.19% (49 857 of 805 206; 95% CI: 6.14%, 6.24%). If there had been only single reading (the first reader decision only) then the recall rate would have been 4.76% (38 295 of 805 206; 95% CI: 4.71%, 4.80%), $P < .001$).

There were 7055 cancers detected by the system of double reading plus arbitration. If there had been only single reading (the first reader decision only), then fewer cancers (n = 6425) would have been detected ($P < .001$, test of two proportions). The second reader detected an additional 627 cancers that were not detected by the first reader.

The additional cancers detected by the second reader (which were not detected by the first reader) were less likely to contain invasive disease (69.5% [417 of 600] were invasive vs 78.0% [4770 of 6114]; $P < .001$). Where invasive disease was present, it was likely to be lower grade in the additional cancers detected by the second reader only, with 32.3% (134 of 415) low grade in comparison to 24.7% (1167 of 4729) and 14.9% (62 of 415) high grade in comparison to 20.9% (986 of 4729) ($P < .001$). The mean pathologic size of the invasive tumors was smaller for the additional cancers detected by the second reader, at 14.2 mm ± 10.6 (n = 416), compared with 16.7 mm ± 12.2 (n = 4761) for cancers detected by the first reader ($P < .001$). There were fewer nodes involved in invasive cancers detected by the second reader only; of these, only 3.2% (13 of 406) had three or more involved nodes, and 12.6% (51 of 406) had one to two involved nodes, compared with 6.2% (291 of 4688) with three or more involved nodes and 17.8% (833 of 4688) with one to two involved nodes in the invasive cancers detected by the first reader ($P < .001$). Figures 2 and 3 show example cases of small invasive cancers and DCIS detected by the second reader only.

For cancers where DCIS only was present (no invasive disease), DCIS grade was lower in cancers detected by the second reader only compared with those cancers detected by the first reader, with 56.9% (87 of 153) high grade compared with 63.9% (743 of 1163) and 17.0% (26 of 153) low grade in comparison to 8.9% (104 of 1163), (P = .007). The results of the sensitivity analysis assuming all missing data were extreme cases did not alter the overall results.

We have follow-up data for 247 885 women from 10 of the first centers to complete the trial. A total of 210 525 women attended a follow-up screening examination. The cancer detection rate overall for the follow-up appointment was 9.7 per 1000 women screened (2043 of 210 525). The cancer detection rate for the follow-up appointment was different in the following three groups: recall by the first reader but not by the second reader and arbitration at the current screening (16 per 1000 women screened [45 of 2800]), recall by the second reader but not by the first reader and arbitration at the current screening (24 per 1000 women screened [47 of 1954]) and all other women who were not recalled by either reader at the current screening (9.3 per 1000 women screened [1839 of 198 602]; $P < .001$). The 3-year interval cancer rate was 2.1 per 1000 women screened (512 of 247 885). The interval cancer rate was different in the following three groups: recall by the first reader but not by the second reader or arbitration at the current screening (5.5 per 1000 women screened [18 of 3281]), recall by the second reader but not by the first reader or arbitration at the current screening (6.1 per 1000 women screened [14 of 2281]), and all other women who were not recalled by either reader at the current screening (1.9 per 1000 women screened [443 of 231 937]; $P < .001$). If we were to assume all of the subsequent excess cancers (interval and 3-year follow-up) in women who had a discordant reading and were not recalled by arbitration were missed, and these same rates applied to the whole data set, then there would have been 752 (10.3%) cancers detected by the second reader only, in

Figure 1

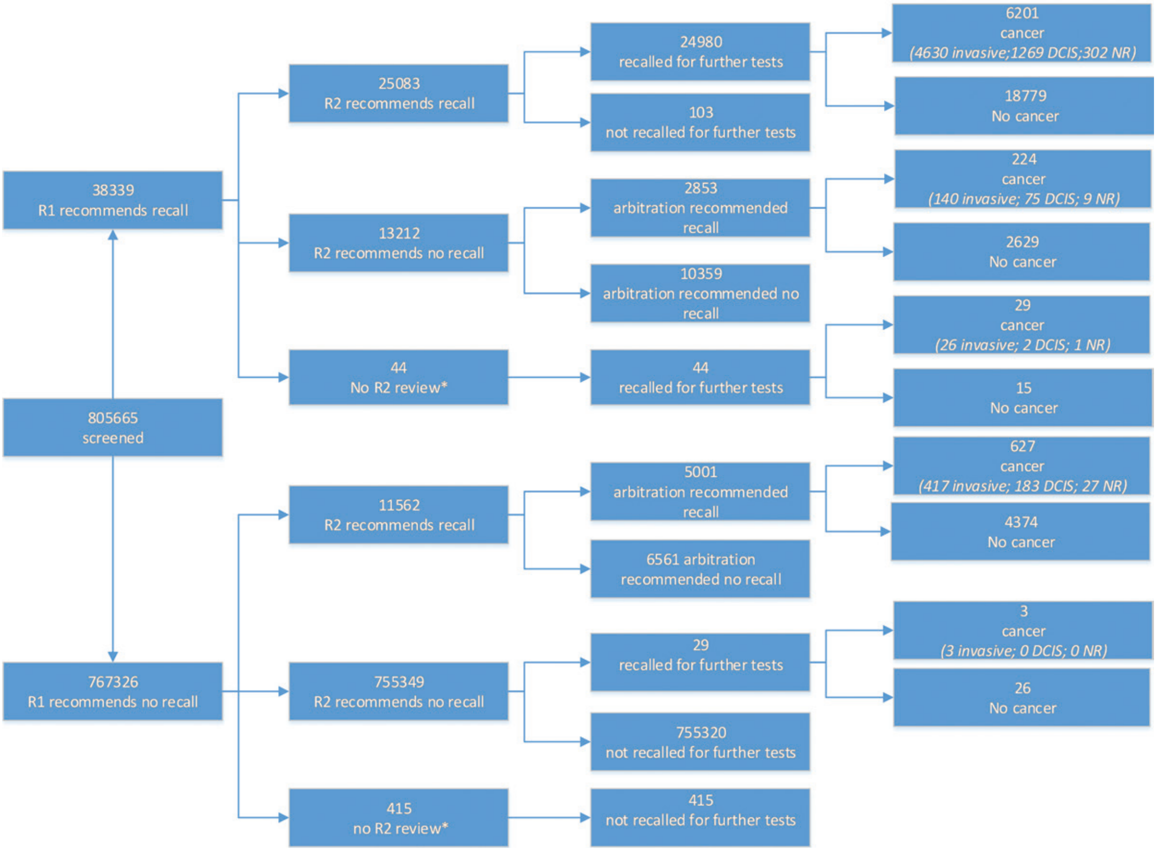
805665
screened
38339
R1 recommends recall
767326
R1 recommends no recall
25083
R2 recommends recall
13212
R2 recommends no recall
44
No R2 review*
11562
R2 recommends recall
755349
R2 recommends no recall
415
no R2 review*
24980
recalled for further tests
103
not recalled for further tests
2853
arbitration recommended recall
10359
arbitration recommended no recall
44
recalled for further tests
5001
arbitration recommended recall
6561 arbitration recommended no recall
29
recalled for further tests
755320
not recalled for further tests
415
not recalled for further tests
6201
cancer
(4630 invasive;1269 DCIS;302 NR)
18779
No cancer
224
cancer
(140 invasive; 75 DCIS; 9 NR)
2629
No cancer
29
cancer
(26 invasive; 2 DCIS; 1 NR)
15
No cancer
627
cancer
(417 invasive; 183 DCIS; 27 NR)
4374
No cancer
3
cancer
(3 invasive; 0 DCIS; 0 NR)
26
No cancer

Figure 1: Charts show flow of women through the study. * = Women whose mammograms were reviewed only by a single reader were excluded from further analysis. *DCIS* = ductal carcinoma in situ, *NR* = not recorded, *R1* = first reader, *R2* = second reader.

addition to 6536 detected by the first reader.

Discussion

In this large population-based cohort study nested within a trial we found that the addition of a second reader to interpret breast screening mammograms, plus arbitration of discordant examinations, reduced recall rate and increased cancer detection rate. The second reader detected an extra 627 (of 7055 [8.9%]) cancers not detected by the first reader, but these were smaller and lower grade and were less likely to be invasive or have involved nodes. These characteristics are indicative of earlier detection and a potential benefit from less aggressive, more successful treatment, but are also suggestive of overdiagnosis of disease. While overdiagnosis is more associated with smaller, lower grade, noninvasive disease without involved nodes, we cannot accurately predict which individual cancers will develop symptomatically.

Previous studies using digital mammography found higher recall rates using double reading (4.8%–4.9%) than single reading (4.6%) (23,24). A recent analysis (23) suggested that double reading may not be cost effective. We found that with effective arbitration of discordant examinations, a second reader can reduce recall rates, but a formal cost-benefit analysis would be needed to assess the incremental benefit of the time involved in the second round of interpretations in the optimal strategy. Previous digital mammography studies have been small and have revealed no statistically significant difference in cancer detection rates or the size, grade, and type of cancer between single and double reading (22–24). Our study is an order of magnitude larger than these studies and indicates that the addition of a second reader increases cancer detection rates, although the additional cancers detected are smaller, and of a lower grade and stage. The inconsistencies observed between our study and previous studies may reflect the greater statistical power of our study to detect small differences.

Policy makers routinely evaluate how to deliver breast screening in optimal ways. In France, a recommendation has been made to expand the use of second readers on the basis of increased cancer detection (from older film mammography studies) and quality assurance (30). Conversely, Spanish researchers have

Recall Rates and Characteristics of Cancers Detected in 805 206 Women for the System Implemented in the United Kingdom

Parameter	Double Reading and Arbitration	First Reader Recall	Second Reader Recall Only, not First Reader	*P* Value
No. of women recalled	32 863 (4.08)	38 295 (4.76)	11 562 (1.44)	…
No. of benign biopsies	8369 (1.04)	7192 (0.89)	1167 (0.14)	…
No. of cancers detected	7055 (0.88)	6425 (0.80)	627 (0.08)	…
Invasive disease present?				
Yes	5190 (77.3)	4770 (78.0)	417 (69.5)	
No	1527 (22.7)	1344 (22.0)	183 (30.5)	<.001
Not recorded	338	311	27	
Invasive disease grade				.
Grade 3	1048 (20.4)	986 (20.9)	62 (14.9)	.
Grade 2	2797 (54.3)	2576 (54.5)	219 (52.8)	<.001
Grade 1	1302 (25.3)	1167 (24.7)	134 (32.3)	
Not recorded	43	41	2	
No. of positive axillary nodes in women with invasive disease				
0	3909 (76.7)	3564 (76.0)	342 (84.2)	
1–2	884 (17.3)	833 (17.8)	51 (12.6)	<.001
3+	304 (6.0)	291 (6.2)	13 (3.2)	
Not recorded	93	82	11	
Characteristics of invasive cancers				
No.	5180	4761	416	
Mean size (mm) ± standard deviation	16.5 ± 12.1	16.7 ± 12.2	14.2 ± 10.6	<.001
Median size (mm)	14	14	11	
Interquartile range for size (mm)	9–20	9–20	8–17	
DCIS grade				
High	830 (63.1)	743 (63.9)	87 (56.9)	
Intermediate	356 (27.1)	316 (27.2)	40 (26.1)	.007
Low	130 (9.9)	104 (8.9)	26 (17.0)	
None	2	2	0	
Not recorded	209	179	30	

Note.—Unless otherwise noted, data are numbers of women, with percentages in parentheses. The system is double reading plus arbitration of disagreements. The table shows results of comparison of the characteristics of all cancers detected by the first reader with those of the extra cancers detected by the second reader alone. DCIS = ductal carcinoma in situ.

suggested that double reading may not be cost effective (on the basis of results of small digital mammography studies suggesting increased recall rates) (23). Our findings indicate an increase in cancer detection with a second reader using digital mammography and that recall rates can be reduced with effective arbitration. To fully understand the difference in outcomes between screening programs using single or double reading requires a randomized controlled trial. Future research may also investigate the effect of a second reader when using breast tomosynthesis.

This study had limitations. First, some women recalled by one reader received only a reference standard of arbitration, and only those recalled by arbitration received further testing (eg, diagnostic biopsy). It is possible that some women not recalled by arbitration did have cancer that was not detected by this reference standard. In a study that predominantly used film mammography, Hofvind and colleagues (31) reported that the rate of interval cancers was higher among women who had discordant interpretations of their mammograms (ie, where one reader recommended recall and the other did not) and who were not recalled than among the whole screening population (2.9 per 1000 vs 1.7 per 1000). Similarly, in our study, interval cancer rates were higher in women whose cases were arbitrated and not recalled at the current round (6.1 per 1000 women in those recalled by reader 2 only and 5.5 per 1000 women in those recalled by reader 1 only) than in other women not recalled at the current round (1.9 per 1000 women screened). Ascertainment of interval cancers is unlikely to be complete (particularly from year 3) because of delays in data transfer from the English cancer registries to screening units. For the same groups of women whose cases were arbitrated and not recalled at the current round, we found a similar excess in cancer detection rates at the subsequent screening round. This excess may be due to a combination of cancers missed at the current screening round by arbitration and cases with a

Figure 2

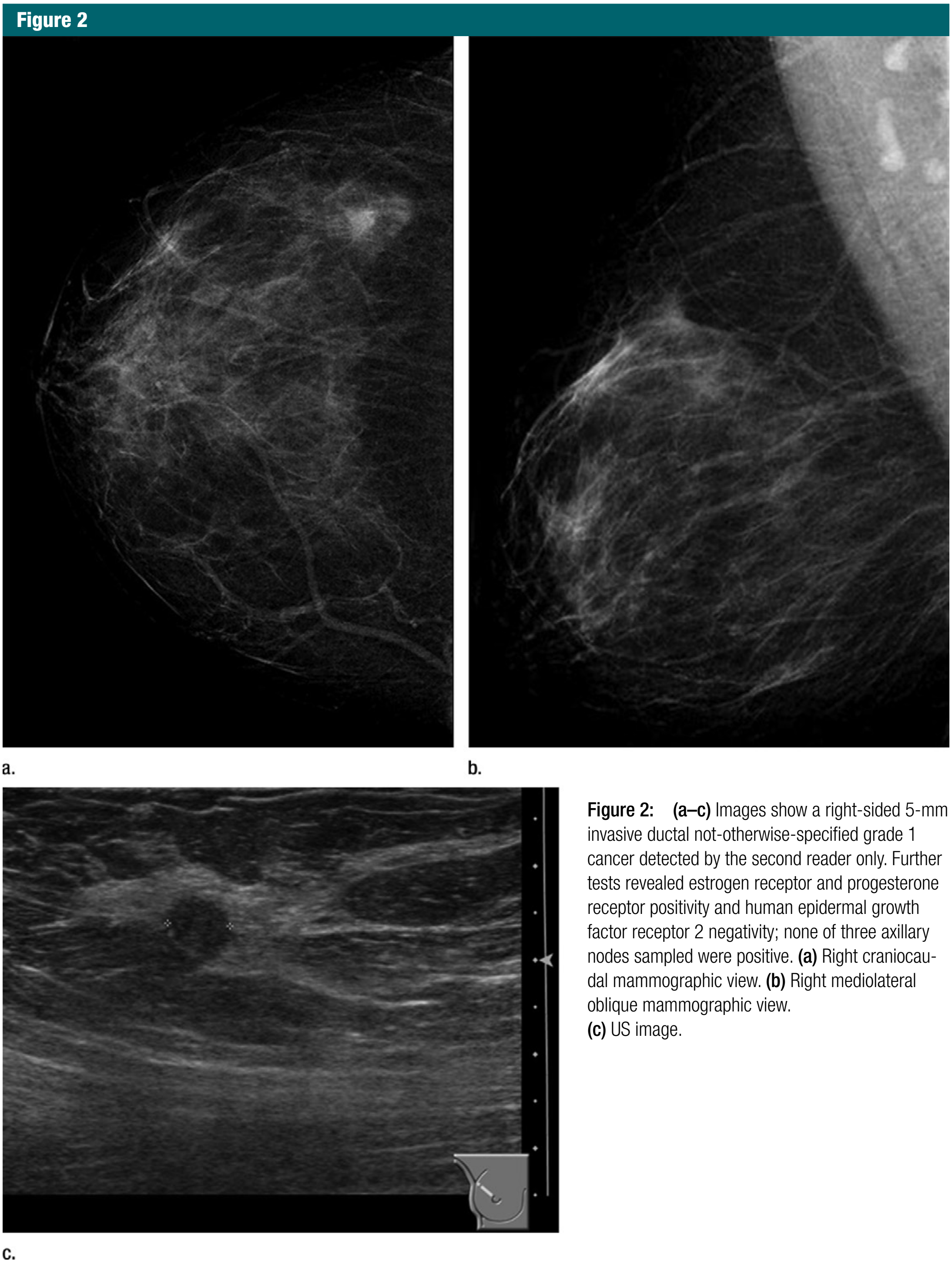

Figure 2: **(a–c)** Images show a right-sided 5-mm invasive ductal not-otherwise-specified grade 1 cancer detected by the second reader only. Further tests revealed estrogen receptor and progesterone receptor positivity and human epidermal growth factor receptor 2 negativity; none of three axillary nodes sampled were positive. **(a)** Right craniocaudal mammographic view. **(b)** Right mediolateral oblique mammographic view. **(c)** US image.

discordant reading having other characteristics (eg, increased breast density) associated with a higher risk of cancers developing between screening rounds. If we assumed the excess was purely due to cancers missed by arbitration, then 10.3% of cancers would have been detected by reader 2 only. This is higher than the 8.9% we report, but it may be an overestimate because of the inclusion of cancers that have developed in the screening interval. Second, we assumed that the actions of the first reader would be the same as the actions of a single reader. Readers working alone may

Figure 3

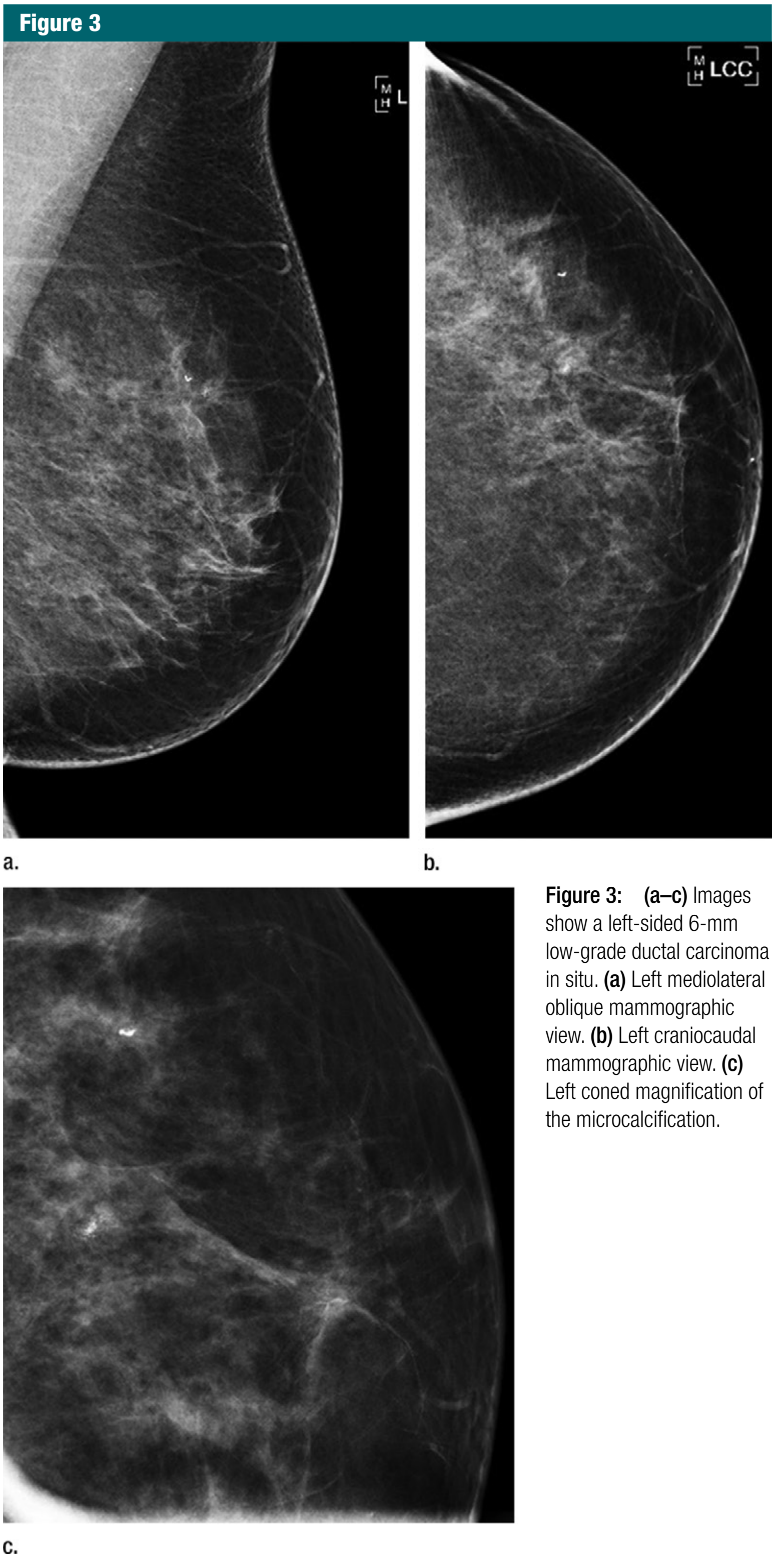

Figure 3: **(a–c)** Images show a left-sided 6-mm low-grade ductal carcinoma in situ. **(a)** Left mediolateral oblique mammographic view. **(b)** Left craniocaudal mammographic view. **(c)** Left coned magnification of the microcalcification.

operate at a different standard or recall threshold if there is no second reader to pick up missed cancers and no arbitration to reduce false-positive recalls. Third, although performing this study in a trial setting minimized missing data, there remained some missing information about cancer characteristics. The results of the sensitivity analysis assuming that all missing data were extreme cases did not alter the overall results. Finally, while readers independently examined mammograms, they could access the decision of the first reader by examining notes. This is not a normal part of reading in a busy population screening program, but if it occurred would support our null hypothesis and underestimate the incremental value of a second reader (if second readers were aligning their results with that of the first reader, cancers detected by the second reader would not have different characteristics from those detected by the first reader).

In conclusion, in this large population-based cohort study, the use of a second reader plus arbitration in mammography reduced recall rates and improved cancer detection. The extra cancers detected were smaller and lower grade and were less likely to be invasive or have involved nodes. Detecting these extra cancers may be associated with detecting important pathologic findings earlier, but it may also be associated with increased overdiagnosis from screening. Further analysis of follow-up data on outcomes is required to understand the balance of the benefits and harms of detecting these extra cancers. Policy makers should consider the overall harms and benefits when deciding whether to use a second reader, bearing in mind that a single reader might not perform the same way a first reader working as part of a team might.

Disclosures of Conflicts of Interest: **S.T.** disclosed no relevant relationships. **D.J.** disclosed no relevant relationships. **C.S.** disclosed no relevant relationships. **M.G.W.** disclosed no relevant relationships. **J.D.** disclosed no relevant relationships. **A.C.** disclosed no relevant relationships.

References

1. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA Cancer J Clin 2015;65(2):87–108.

2. Tabár L, Vitak B, Chen TH, et al. Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades. Radiology 2011;260(3):658–663.

3. Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. The benefits and harms of breast cancer screening: an independent review. Br J Cancer 2013;108(11):2205–2240.

4. Wilson R, Liston J. Quality Assurance Guidelines for Breast Cancer Screening Radiology: NHS Breast Screening Programme Publication Number 59. Sheffield, England: NHS Cancer Screening Programmes, 2011.

5. Perry N, Broeders M, de Wolf C, Törnberg S, Holland R, von Karsa L. European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition—summary document. In: Perry N, Broeders M, de Wolf C, Törnberg S, Holland R, von Karsa L, eds. European guidelines for quality assurance in breast cancer screening and diagnosis. 4th ed. Luxembourg: European Commission, Office for Official Publications of the European Union, 2013; XIV–XX.

6. Lehman CD, Wellman RD, Buist DS, et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Intern Med 2015;175(11):1828–1837.

7. Anderson ED, Muir BB, Walsh JS, Kirkpatrick AE. The efficacy of double reading mammograms in breast screening. Clin Radiol 1994;49(4):248–251.

8. Anttinen I, Pamilo M, Soiva M, Roiha M. Double reading of mammography screening films: one radiologist or two? Clin Radiol 1993;48(6):414–421.

9. Harvey SC, Geller B, Oppenheimer RG, Pinet M, Riddell L, Garra B. Increase in cancer detection and recall rates with independent double interpretation of screening mammography. AJR Am J Roentgenol 2003;180(5):1461–1467.

10. Brown J, Bryan S, Warren R. Mammography screening: an incremental cost effectiveness analysis of double versus single reading of mammograms. BMJ 1996;312(7034):809–812.

11. Ciatto S, Ambrogetti D, Bonardi R, et al. Second reading of screening mammograms increases cancer detection and recall rates: results in the Florence screening programme. J Med Screen 2005;12(2):103–106.

12. Taylor P, Potts HW. Computer aids and human second reading as interventions in screening mammography: two systematic reviews to compare effects on cancer detection and recall rate. Eur J Cancer 2008;44(6):798–807.

13. Warren RM, Duffy SW. Comparison of single reading with double reading of mammograms, and change in effectiveness with experience. Br J Radiol 1995;68(813):958–962.

14. Georgian-Smith D, Moore RH, Halpern E, et al. Blinded comparison of computer-aided detection with human second reading in screening mammography. AJR Am J Roentgenol 2007;189(5):1135–1141.

15. Leivo T, Salminen T, Sintonen H, et al. Incremental cost-effectiveness of double-reading mammograms. Breast Cancer Res Treat 1999;54(3):261–267.

16. Blanks RG, Wallis MG, Moss SM. A comparison of cancer detection rates achieved by breast cancer screening programmes by number of readers, for one and two view mammography: results from the UK National Health Service breast screening programme. J Med Screen 1998;5(4):195–201.

17. Yen MF, Tabár L, Vitak B, Smith RA, Chen HH, Duffy SW. Quantifying the potential problem of overdiagnosis of ductal carcinoma in situ in breast cancer screening. Eur J Cancer 2003;39(12):1746–1754.

18. van Luijt PA, Heijnsdijk EA, Fracheboud J, et al. The distribution of ductal carcinoma in situ (DCIS) grade in 4232 women and its impact on overdiagnosis in breast cancer screening. Breast Cancer Res 2016;18(1):47.

19. Thurfjell E. Mammography screening methods and diagnostic results. Acta Radiol Suppl 1995;395:1–22.

20. Screening and Immunisations Team, Health and Social Care Information Centre. Breast Screening Programme, England—2012–13. Leeds, England: Health and Social Care Information Centre, 2014.

21. U.S. Food and Drug Administration. Mammography Quality Standards Act and Program. Silver Spring, Md: U.S. Food and Drug Administration, 2017.

22. Houssami N, Macaskill P, Bernardi D, et al. Breast screening using 2D-mammography or integrating digital breast tomosynthesis (3D-mammography) for single-reading or double-reading: evidence to guide future screening strategies. Eur J Cancer 2014;50(10):1799–1807.

23. Posso M, Carles M, Rué M, Puig T, Bonfill X. Cost-effectiveness of double reading versus single reading of mammograms in a breast cancer screening programme. PLoS One 2016;11(7):e0159806.

24. Posso MC, Puig T, Quintana MJ, Solà-Roca J, Bonfill X. Double versus single reading of mammograms in a breast cancer screening programme: a cost-consequence analysis. Eur Radiol 2016;26(9):3262–3271.

25. Posso M, Puig T, Carles M, Rué M, Canelo-Aybar C, Bonfill X. Effectiveness and cost-effectiveness of double reading in digital mammography screening: a systematic review and meta-analysis. Eur J Radiol 2017;96(Supplement C):40–49.

26. Taylor-Phillips S, Wallis MG, Jenkinson D, et al. effect of using the same vs different order for second readings of screening mammograms on rates of breast cancer detection: a randomized clinical trial. JAMA 2016;315(18):1956–1965.

27. Taylor-Phillips S, Wallis MG, Gale AG. Should previous mammograms be digitised in the transition to digital mammography? Eur Radiol 2009;19(8):1890–1896.

28. Borrelli C, Cohen S, Duncan A, et al. NHS Breast Screening Programme Clinical guidance for breast cancer screening assessment: NHSBSP publication number 49. 4th ed. London, England: Public Health England, 2016.

29. R Development Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2017.

30. Cases C, Di Palma M, Drahl E, et al. Améliorons le dépistage du cancer du sein: concertation citoyenne et scientifique. Rapport du comité d'orientation. http://www.concertation-depistage.fr/wp-content/uploads/2016/10/depistage-cancer-sein-rapport-concertation-sept-2016.pdf. September 2016. Accessed May 2017.

31. Hofvind S, Geller BM, Rosenberg RD, Skaane P. Screening-detected breast cancers: discordant independent double reading in a population-based screening program. Radiology 2009;253(3):652–660.