Effect of Using the Same vs Different Order for Second Readings of Screening Mammograms on Rates of Breast Cancer Detection A Randomized Clinical Trial

Taylor-Phillips, S. , Wallace, M. G. , Jenkinson, D. , Adekanmbi, V. , Parsons, H. , Dunn, J. , Stallard, N. , Szczepura, A. , Gates, S. , Kearins, O. , Duncan, A. , Hudson, S. and Clarke, A.

Original citation:
Taylor-Phillips, S. , Wallace, M. G. , Jenkinson, D. , Adekanmbi, V. , Parsons, H. , Dunn, J. , Stallard, N. , Szczepura, A. , Gates, S. , Kearins, O. , Duncan, A. , Hudson, S. and Clarke, A. (2016) Effect of Using the Same vs Different Order for Second Readings of Screening Mammograms on Rates of Breast Cancer Detection A Randomized Clinical Trial. The Journal of the American Medical Association, volume 215 (18): 1956-1965
http://dx.doi.org/10.1001/jama.2016.5257

DOI: 10.1001/jama.2016.5257
ISSN: 0098-7484
ESSN: 1538-3598

Publisher: American Medical Association

Original Investigation

Effect of Using the Same vs Different Order for Second Readings of Screening Mammograms on Rates of Breast Cancer Detection
A Randomized Clinical Trial

Sian Taylor-Phillips, PhD; Matthew G. Wallis, FRCR; David Jenkinson, PhD; Victor Adekanmbi, PhD; Helen Parsons, PhD; Janet Dunn, PhD; Nigel Stallard, PhD; Ala Szczepura, DPhil; Simon Gates, PhD; Olive Kearins, MSc; Alison Duncan, FRCR; Sue Hudson, MSc; Aileen Clarke, MD

IMPORTANCE Interpreting screening mammograms is a difficult repetitive task that can result in missed cancers and false-positive recalls. In the United Kingdom, 2 film readers independently evaluate each mammogram to search for signs of cancer and examine digital mammograms in batches. However, a vigilance decrement (reduced detection rate with time on task) has been observed in similar settings.

OBJECTIVE To determine the effect of changing the order for the second film reader of batches of screening mammograms on rates of breast cancer detection.

DESIGN, SETTING, AND PARTICIPANTS A multicenter, double-blind, cluster randomized clinical trial conducted at 46 specialized breast screening centers from the National Health Service Breast Screening Program in England for 1 year (all between December 20, 2012, and November 3, 2014). Three hundred sixty readers participated (mean, 7.8 readers per center)—186 radiologists, 143 radiography advanced practitioners, and 31 breast clinicians, all fully qualified to report mammograms in the NHS breast screening program.

INTERVENTIONS The 2 readers examined each batch of digital mammograms in the same order in the control group and in the opposite order to one another in the intervention group.

MAIN OUTCOMES AND MEASURES The primary outcome was cancer detection rate; secondary outcomes were rates of recall and disagreements between readers.

RESULTS Among 1 194 147 women (mean age, 59.3; SD, 7.49) who had screening mammograms (596 642 in the intervention group; 597 505 in the control group), the images were interpreted in 37 688 batches (median batch size, 35; interquartile range [IQR]; 16-46), with each reader interpreting a median of 176 batches (IQR, 96-278). After completion of all subsequent diagnostic tests, a total of 10 484 cases (0.88%) of breast cancer were detected. There was no significant difference in cancer detection rate with 5272 cancers (0.88%) detected in the intervention group vs 5212 cancers (0.87%) detected in the control group (difference, 0.01% points; 95% CI, −0.02% to 0.04% points; recall rate, 24 681 [4.14%] vs 24 894 [4.17%]; difference, −0.03% points; 95% CI, −0.10% to 0.04% points; or rate of reader disagreements, 20 471 [3.43%] vs 20 793 [3.48%]; difference, −0.05% points; 95% CI, −0.11% to 0.02% points).

CONCLUSIONS AND RELEVANCE Interpretation of batches of mammograms by qualified screening mammography readers using a different order vs the same order for the second reading resulted in no significant difference in rates of detection of breast cancer.

TRIAL REGISTRATION isrctn.org Identifier: ISRCTN46603370

Editorial page 1951

Supplemental content at jama.com

CME Quiz at jamanetworkcme.com and CME Questions page 2013

Author Affiliations: Warwick Medical School, the University of Warwick, Coventry, United Kingdom (Taylor-Phillips, Jenkinson, Adekanmbi, Parsons, Dunn, Stallard, Szczepura, Gates, Clarke); Cambridge Breast Unit, Cambridge Universities NHS Foundation Trust and NIHR Cambridge Biomedical Research Centre, Cambridge, United Kingdom (Wallis); Screening QA Service (Midland & East), Seaton House, City Link, Nottingham, United Kingdom (Kearins); Warwickshire, Solihull and Coventry Breast Screening Service, University Hospitals Coventry and Warwickshire, Coventry, United Kingdom (Duncan); Peel & Schriek Consulting Limited, London, United Kingdom (Hudson).

Corresponding Author: Sian Taylor-Phillips, PhD, University of Warwick, Warwick Medical School, Gibbett Hill Road, Coventry, Warwickshire CV47AL, United Kingdom (s.taylor-phillips@warwick.ac.uk).

JAMA. 2016;315(18):1956-1965. doi:10.1001/jama.2016.5257

Breast cancer screening detects 8.6 cancers per thousand women screened triennially (equivalent to 18 000 cancers per year) in the United Kingdom,[1] and 4.2 cancers per thousand women screened annually in the United States.[2] However, another 2.9 cancers per thousand women screened in the United Kingdom[3] (equivalent to 6030 cancers per year)[1] and 0.9 cancers per thousand women screened in the United States are detected between screening rounds in screened women.[2] These arise through cancers growing between screening rounds, and cancers missed at screening. An additional 3.3% of women in the United Kingdom (69 700 each year)[1] and 9.3% of women in the United States[2] experience false-positive recalls at each screening round.

Interpreting screening mammograms is a difficult and repetitive visual search task, for which characteristics of cancer are disguised among background breast parenchyma resulting in false-positive recalls and missed cancers. In similar visual search tasks, a vigilance decrement of decreasing detection rates with time on task has been observed in a large number of psychological laboratory experiments,[4,5] for example, assembly line inspection tasks,[6] airport baggage screening,[7] driving,[8] piloting airplanes,[9] and operating military drones.[10] An effect similar to the vigilance decrement has been observed when examining tests sets of x-rays including mammograms in laboratory conditions although the phenomenon has not previously been explored in breast screening practice.[11,12]

In the United Kingdom, 2 film readers independently examine each woman's mammograms for signs of cancer. In this study, we investigated whether a vigilance decrement to detect cancer in breast screening practice exists and whether changing the order in which the 2 experts examined the batch of mammograms could increase the cancer detection rate, through readers' experiencing peak vigilance at differing points within the reading batch when examining different women's mammograms.

Methods

Study Design

The Changing Case Order to Optimise Patterns of Performance in Screening (CO-OPS), a pragmatic, double-blind, cluster randomized clinical trial, measured whether a vigilance decrement in breast cancer screening exists and whether changing case order can increase cancer detection.

Ethical approval was granted by the Coventry and Warwickshire National Health Service (NHS) Research Ethics Committee on June 27, 2012[12] (WM/0182). Each director of breast screening provided written informed consent. The trial protocol is provided in Supplement 1 and published elsewhere.[13] The statistical analysis plan was finalized before any data were collected (Supplement 2) .

Intervention and Outcomes

The study compared 2 parallel groups, each split into 2 subgroups to ensure blinding of the readers. The intervention group involved the 2 readers reviewing the batch in the opposite order of each other; 1 forward, 1 in reverse. Hence, the 2 subgroups: first reader forward, second reader reverse, and first reader reverse, second reader forward. The control group required the readers to read the batch in the same order as each other; the subgroups being either both forward (which is current practice) or both in reverse (to maintain the blinding of a reader to trial group because the readers would be aware that they were reading a batch in reverse). Thus, each batch (cluster) was randomized with equal probability to 1 of 4 groups.

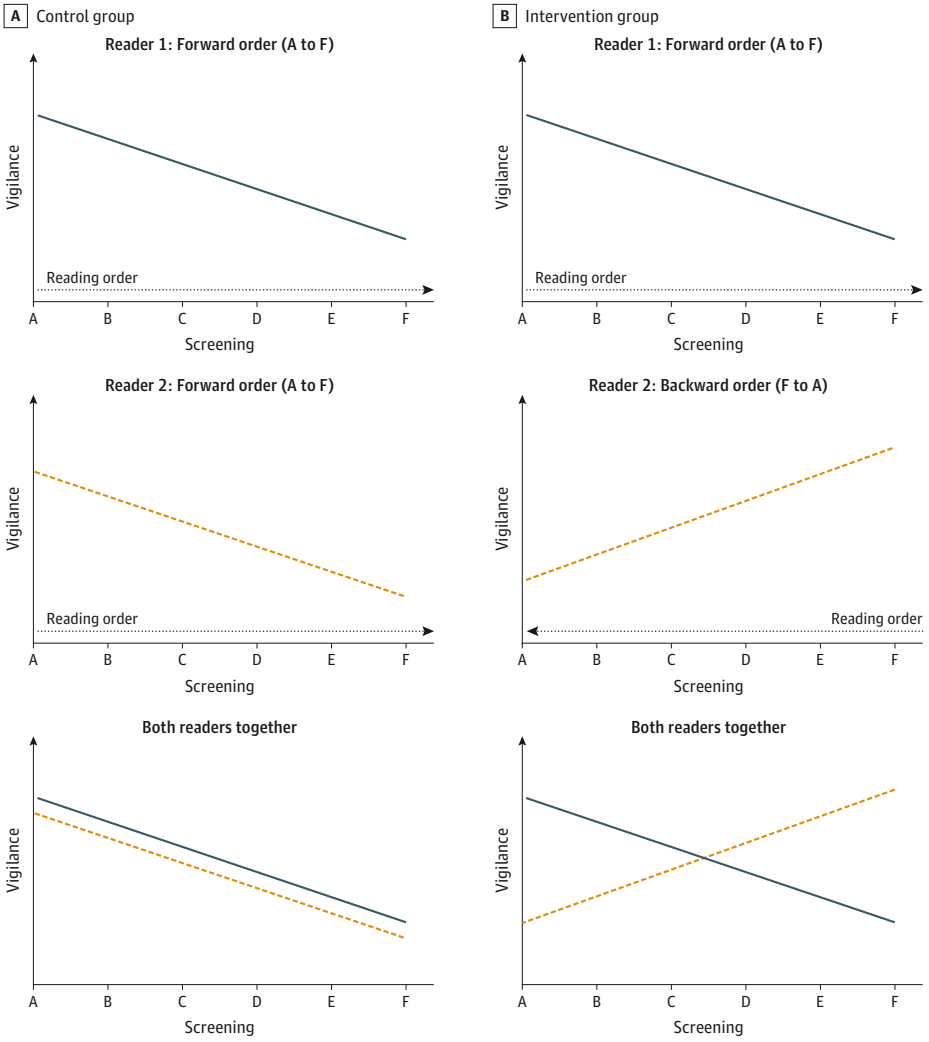
The primary outcome was cancer detection rate, (number of women with cancer detected as a proportion of all women screened) because this is the clinically relevant outcome of interest. Secondary outcomes of recall rate (secondary outcome 1) and rate of disagreement between the readers (secondary outcome 2) are designed to examine the proposed mechanism of action. The idea is that reversing the order for 1 reader results in high-vigilance states occurring for the 2 readers when examining different women's mammograms, so the cancers are detected by at least 1 of the experts, as outlined in **Figure 1**. If a reader in a high-vigilance state detected a cancer missed by his/her colleague in a low-vigilance state, then this would lead to a disagreement between them. All disagreements are arbitrated either by a third reader or group of readers for the final decision of whether to recall the woman for further tests. Assuming the arbitration process performs better than random chance the increases in disagreements would lead to increases in recall rate and cancer detection rate.

Participants

Centers were recruited at radiology meetings, through local radiology, radiography, and quality assurance groups and through direct telephone and email contact. The study comprised 46 breast screening centers using digital mammography, each consisting of groups of between 1 and 3 hospitals sharing the same computer system for storing women's health records. Characteristics of breast screening centers in England that participated in the trial compared with those that did not is provided in eTable 1 in Supplement 2. The trial ran for 1 year at each center, with individual centers starting the study when local consent and research and development approvals were obtained, (start dates were all between December 20, 2012, and November 3, 2014). One center completed only 4 months of the study due to local technical and workforce issues.

Mammograms from women attending routine breast cancer screening at these centers during the study period were included. These were arranged into batches of approximately 40 women pursuant to standard practice in the United Kingdom. All mammograms taken during the study period were included in the trial, regardless of when they were examined. Each batch contained all cases from a single mammography acquisition machine in a single day. Informed consent was at the center level, with consent of individual women considered impractical for this system-level intervention. In the United Kingdom, women aged 50 to 70 years are invited to breast screening every 3 years. This study also includes women aged 47 through 49 years and 71 through 73 years who were participating in the age extension trial (NCT01081288), and a small pro-

Figure 1. Proposed Mechanism of Action of Changing Case Order Intervention, Assuming the Hypothesized Vigilance Decrement



Each screening represents examining a set of 4 mammograms, mediolateral oblique and craniocaudal views of both breasts for one woman.

portion of older women (2.3% of women in the trial) who self-refer as part of the program. Women who presented to clinics symptomatically and for high-familial risk were excluded.

All readers undergo formal training and are accredited by the NHS Breast Screening Programme. They are required to read a minimum of 5000 cases per year, participate in assessment clinics, formally audit their own performance against their peers, and maintain ongoing professional development including participating annually in the Personal Performance in Mammographic Screening test set.[14] Each center annually measures and reports results against targets including recall rate, cancer detection rate, and small-cancer detection rate, and continuously audits performance through monthly review of interval cancers diagnosed symptomatically between screening rounds,[14] and monthly checks of mammography acquisition and display equipment and reading room background light levels.[15] Each woman's mammograms are examined by 2 readers located in the same breast screening center. Readers are instructed to examine the batches independently, but can access the other reader's decision by opening the patient records. In 16 of the 46 centers, workflow systems were designed to blind reader 2 to the decision of reader 1. All centers used arbitration when the 2 readers disagreed, with 13 centers using a single third reader, and 33 centers using group consensus of 2 or more readers.

Randomization and Blinding

The randomization took place immediately prior to opening each batch for examination using the Intersystems Caché $RANDOM function within the computer software that the UK National Breast Screening Service (NBSS) uses to manage the work. After randomization the software automatically displayed the cases in the chosen order to the first and second reader. Readers were aware of the reading order but were blinded to trial group. The trial statistician and the women screened were also blinded to trial group. The unit of randomization was a batch of mammograms, whereas the unit of observation was the individual mammogram. Simple random-

ization was used without stratification or minimization due to the large number of clusters randomized.

Data Collection

The data were collected via an adaptation to the NBSS computer system, which created new tables within the software to record data items pertaining to the trial. The outcomes for every woman screened (including both readers' decision, time of decision, and results of all follow-up tests including biopsy) were added to NBSS as part of each center's annual reporting requirements to reduce missing data. The data were extracted through NBSS from each center, exporting data in Excel format. The data sets from each center were merged using Excel and R (version 3.0.3 in RStudio version 0.98.501). Cancer was defined as needle biopsy or surgery samples that tested positive for ductal carcinoma in situ or invasive cancer. Recall for further tests was taken directly from NBSS, which records this decision to enable the follow-up appointment to be made. Disagreement was defined by examining whether the recommendation of whether to recall differed between the first and second readers.

Sample Size

Prior to the study (year 2011-2012) the breast cancer detection rate in the United Kingdom was 7.8 per thousand women screened.[16] Three years of observational data on patterns of cancer detection with time on task was extracted from routine records at 8 breast screening centers in 1 English region. This suggested that the intervention may result in 1 extra cancer detected per 2000 women screened, an increase to 8.3 per thousand women screened. To detect such an increase required a sample size of 501 361 women in each group, using a 5% significance level and 80% power. The trial had a cluster design, the unit of randomization being the batch, so the sample size needed to be inflated by the design effect. The intercluster correlation coefficient was estimated to be 0.002, resulting in a design effect of 1.09, assuming an average cluster size of 40. Hence, the total sample size required was 1 093 780, which is equivalent to the annual caseload of 44 centers. There were no interim analyses or stopping rules.

Statistical Analysis

We used multivariable multilevel logistic regression to analyze factors associated with breast cancer detection, recall, and disagreement rates due to the hierarchical nature of the data sets. Analysis was intention to treat, with those not receiving the intervention as allocated included in the analysis. However, women lost to follow-up, technical recalls (mammograms were of insufficient quality to read), and second screening of the same woman were excluded. A three-level multilevel model for woman screened (level 1) nested in a batch (level 2) and within a center (level 3) was specified. Four models were constructed for each of the rates as stated above. The first model, a null model without any variable was specified to decompose the amount of variance that existed at each level, the second model included the intervention only, the third model included adjustment for known factors associated with cancer and recall (woman's age and whether she had previously attended screening), whereas the fourth model added the intervention to the adjusted model. All multilevel modeling was performed using MlwiN 2.35[17] called from Stata statistical software for Windows version 14[18] using the runmlwin routine. For the multilevel logistic regression models, (iterative generalized least squares; penalized quasi-likelihood) IGLS PQL2 estimation was used.[19] Two-tailed tests were used, with *P* values <.05 considered significant. The fixed effects (ie, measures of association) are presented as adjusted odds ratios (ORs) with their corresponding 95% CIs. Measures of random effects included intracluster correlation (ICC) and median odds ratio (MOR).[20] The ICC was calculated by the linear threshold according to the formula used by Snijders et al,[21] whereas MOR is a measure of unexplained cluster heterogeneity. Methods used for calculating MOR have been described elsewhere.[20,22] Positive predictive value was also calculated in the intervention and control groups as the proportion of recalled cases in which cancer was detected.

The same models were constructed for 3 predefined subgroups: women younger than 53 years (in whom the intervention may be more effective due to higher breast density increasing the task difficulty); the first and last 5 cases in each batch (in which any difference in vigilance would be at its maximum in the intervention group); and the first batch of the day (to examine whether the effectiveness of the intervention may be masked by examining a number of batches in succession). An exploratory post hoc subgroup analysis of cases, which are not in the first batch of the day for either reader, used the same model structure (to investigate intervention effectiveness when readers may be fatigued).

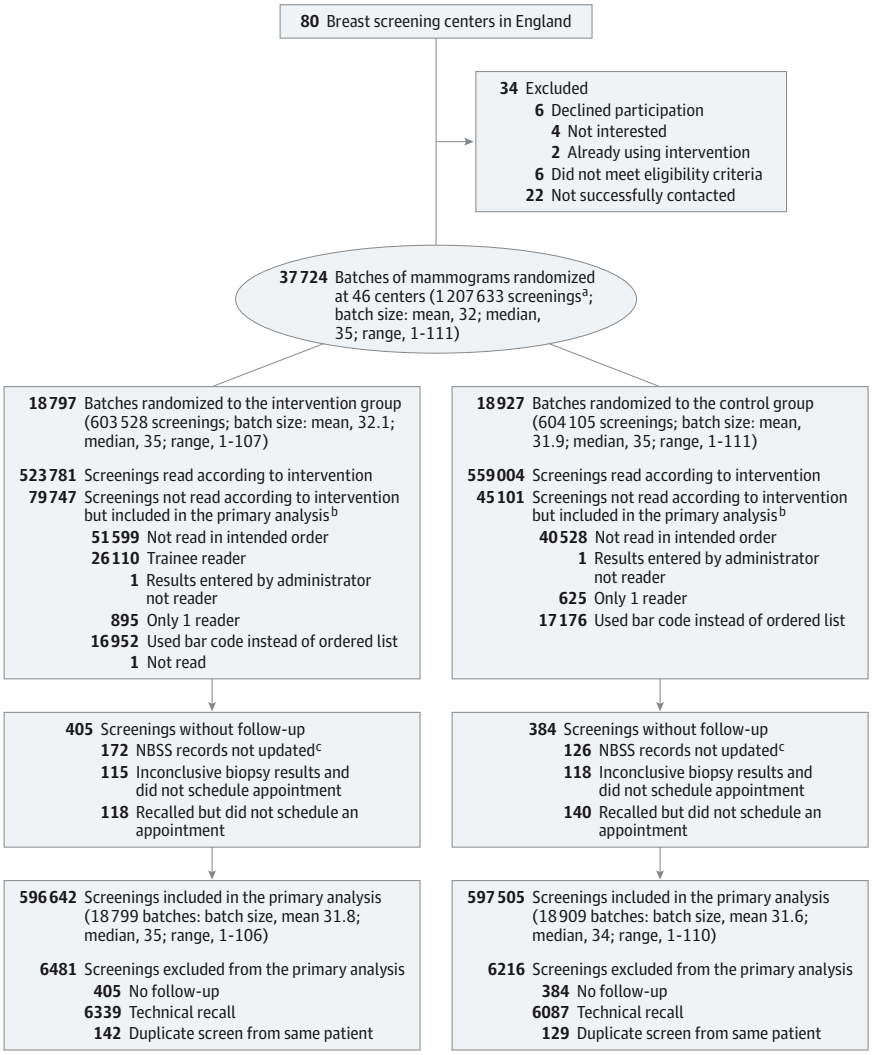
An exploratory post hoc analysis to measure whether there is a vigilance decrement of decreasing sensitivity to detect cancer with time spent on task, the position in the batch (ie, first, second, third...) was added as a variable to the unadjusted and adjusted models of cancer detection outlined above. For this analysis the cancer detection rate outcome was personalized to the individual reader who first examined the case, so the outcome had an additional requirement of being correctly identified by the first reader for recall, as well as having cancer identified on follow-up tests. The same modeling approach was applied to recall rate, to measure any systematic change with time on task. In this case, it was the recall rate for the first reader, rather than overall from the process that was analyzed. Further exploratory post hoc analysis was conducted to determine whether the lack of effect of the intervention was associated with reader 2 not being blinded to the decision of reader 1 at some trial centers. Including only the subgroup of centers in which reader 2 was blinded to the decision of reader 1, cancer detection rates and recall rates in the intervention group were calculated, and compared with those in the control group.

Results

Flow of Women in the Trial

A total of 1 207 633 women were included in the trial (**Figure 2**). There were 3 causes of loss to follow up: 258 (0.02%) were

Figure 2. Study Flow of Trial Comparing Same vs Different Order for Presenting Batches of Mammograms to Breast Screening Readers



[a] Each screening included 4 mammograms (mediolateral oblique and craniocaudal views of both breasts).

[b] For each screening, there may be multiple reasons why they did not receive the allocated intervention.

[c] National Breast Screening Service (NBSS) records are the electronic health records of women screened.

recalled for further tests from screening but did not attend, 233 (0.02%) had an inconclusive needle biopsy test result but refused further tests, and 298 (0.02%) had missing data in the NBSS system. An additional 12 426 cases (1.03%) were judged of insufficient quality for analysis (technical recall) by the first reader so were not read within batch and could not be included in the analysis, and 271 (0.02%) cases were excluded because the same woman had already been screened that year and included in the trial. This occurred primarily when women who had changed primary care practice after moving and consequently were reinvited more quickly than intended.

The intervention and control groups were well matched for baseline characteristics including the age and previous attendance of the women screened and batch length, as detailed in **Table 1**. Mammograms were examined by 360 qualified readers, of which 186 were radiologists, 143 were radiography advanced practitioners, and 31 were breast clinicians. The median batch length was 35 cases (interquartile range [IQR], 16-46). Each reader examined a median of 5640 cases, (IQR, 2599-8458), in a median of 176 batches (IQR, 96-278), including cases in both the intervention and control groups. Between 1 and 26 batches were examined by each reader in a single day (median, 2; IQR, 1-4). Each center examined between 8152 and 72 714 cases (median, 25 540 cases).

Outcomes

The primary outcome, cancer detection rate, was 0.88% (5272 of 596 642) in the intervention group and 0.87% (5212 of 597 505) in the control group (difference, 0.01% points; 95% CI, −0.02% to 0.04%; **Table 2**). The intervention did not affect the cancer detection rate in the unadjusted (OR, 1.01; 95% CI, 0.96-1.06) or adjusted models (OR, 1.01; 95% CI, 0.97-1.06; **Table 3** and eTables 2 and 3 in Supplement 2). In the adjusted model, cancer detection rate increased with each increasing year of age (OR, 1.052; 95% CI, 1.048-1.055) and was higher in women who had not previously attended screening (OR, 1.73; 95% CI, 1.62-1.86). The intervention also had no effect in any of the subgroups of younger age, first and last 5 cases in the

Table 1. Baseline Characteristics for Intervention and Control Groups

	Intervention	Control
Individual level		
Age of women screened, mean (SD), y	59.3 (7.48)	59.3 (7.49)
Had previously attended screening, Total/No. (%)	126 490/596 642 (21.2)	128 217/597 505 (21.5)
Cluster level, median (interquartile range)		
Batch length	35 (16-46)	35 (16-45)
No. of screenings examined by each reader	2848 (1469-4385)	2891 (1543-4458)
No. of batches examined by each reader	86 (52-143)	91 (51-138)
No. of screenings examined at each center	12 496 (8997-16 523)	12 908 (9529-16 418)
No. of batches examined at each center	376 (282-502)	364 (272-521)

Table 2. Primary and Secondary Outcomes in Intervention and Control Groups Overall and by Previous Screening Attendance[a]

	Intervention		Control		
Outcome	No./Total	% (95% CI)	No./Total	% (95% CI)	Difference, % points (95% CI)
Primary Outcome: Cancer Detection Rate					
All screenings	5272/596 642	0.88 (0.86 to 0.91)	5212/597 505	0.87 (0.85 to 0.90)	0.01 (−0.02 to 0.04)
Previous attenders	4214/470 152	0.90 (0.87 to 0.92)	41 22/469 288	0.88 (0.85 to 0.91)	0.02 (−0.02 to 0.06)
Previous nonattenders	1058/126 490	0.84 (0.79 to 0.89)	1090/128 217	0.85 (0.80 to 0.90)	−0.01 (−0.08 to 0.06)
Secondary Outcome: Recall Rate					
All screenings	24 681/596 642	4.14 (4.09 to 4.19)	24 894/597 505	4.17 (4.12 to 4.22)	−0.03 (−0.10 to 0.04)
Previous attenders	14 819/470 152	3.15 (3.10 to 3.20)	14 869/469 288	3.17 (3.12 to 3.22)	−0.02 (−0.09 to 0.05)
Previous nonattenders	9862/126 490	7.80 (7.65 to 7.94)	10 025/128 217	7.82 (7.67 to 7.97)	−0.02 (−0.23 to 0.19)
Secondary Outcome: Disagreement Rate Between Readers					
All screenings	20 471/596 294	3.43 (3.39 to 3.48)	20 793/597 387	3.48 (3.43 to 3.53)	−0.05 (−0.11 to 0.02)
Previous attenders	12 850/469 869	2.73 (2.69 to 2.78)	12 937/469 215	2.76 (2.71 to 2.80)	−0.02 (−0.09 to 0.04)
Previous nonattenders	7621/126 425	6.03 (5.90 to 6.16)	7856/128 172	6.13 (6.00 to 6.26)	−0.10 (−0.29 to 0.08)

[a] Cancer detection rate, recall rate, and rate of disagreement between readers in screenings of previous attenders, screenings of previous nonattenders, and all screenings.

batch, the first batch of the day for both readers, or in batches examined second in the day or later by both readers in either the adjusted or unadjusted models. For batches read first in each workday by both readers, cancer detection rate was 0.83% (580 of 70 071; 95% CI, 0.76%-0.89%) in the intervention group and 0.88% (623 of 70 715; 95% CI, 0.81%-0.95%) in the control group (difference, −0.05% points; 95% CI, −0.15%-0.04% points). For batches read second or subsequent in each workday by both readers, the cancer detection rate was 0.85% (2472 of 289 786; 95% CI, 0.82%-0.89%) in the intervention group and 0.85% (2473 of 290 671; 95% CI, 0.82%-0.88%) in the control group (difference, 0.002% points; 95% CI, −0.045% to 0.050% points).

The intervention did not affect either of the secondary outcomes, recall rate, or rate of disagreements. The recall rate was 4.14% (24 681 of 596 642) in the intervention group and 4.17% (24 894 of 597 505) in the control group (difference, −0.03% points; 95% CI, −0.10% to 0.04% points; Table 2). The rate of disagreement was 3.43% in the intervention group (20 471 of 596 294) and 3.48% (20 793 of 597 387) in the control group (difference, −0.05% points; 95% CI, −0.11% to 0.02% points; Table 2). The intervention had no effect on recall rate in the unadjusted (OR, 0.993; 95% CI, 0.974-1.013) or adjusted (OR, 0.997 95% CI, 0.978-1.016) models (eTable 1 in Supplement 2) or on the rate of disagreement in the unadjusted (OR, 0.994; 95% CI, 0.971-1.019) or adjusted model (OR, 0.997; 95% CI, 0.974-1.020; eTable 4 in Supplement 2). Recall rate was higher with each year of age of the woman screened (OR, 1.008; 95% CI, 1.007-1.010), and was higher in women who had not previously attended breast screening (OR, 2.89; 95% CI, 2.82-2.97). Rate of disagreement was also higher for women at their first screening appointment (OR, 2.17; 95% CI, 2.11-2.24) but lower with each year of increasing age of the woman screened (OR, 0.994; 95% CI, 0.992-0.996). The positive predictive value (PPV) was 21.4% (95% CI, 20.8%-21.9%) in the intervention group and 20.9% (95% CI, 20.4%-21.4%) in the control group (difference, 0.42% points; 95% CI, −0.30% to 1.14% points). The intervention had no effect on any of the subgroups (younger women, first and last cases in the batch, first batch of the day, and second or subsequent batch of the day) for either the adjusted or unadjusted models for either recall rate or rate of disagreements. For batches read first in each workday by both readers, the recall rate was 4.02% (2818 of 70 071; 95% CI, 3.88%-4.17%) in the intervention group and 4.11% (2904 of 70 715, 95% CI, 3.96%-4.25%) in the control group (difference, −0.08% points; 95% CI, −0.29% to 0.12% points), and rate of disagreements was 3.61% (2531 of 70 071; 95% CI, 3.47%-3.75%) in the intervention group and 3.75% (2653 of 70 715; 95% CI, 3.61%-3.89%) in the control group (difference, −0.14% points; 95% CI, −0.34%-0.06% points). For batches read second

or subsequent in each workday by both readers, the recall rate was 4.10% (11 868 of 289 786; 95% CI, 4.02%-4.17%) in the intervention group and 4.15% (12 068 of 290 671; 95% CI, 4.08%-4.22%) in the control group (difference, −0.06% points; 95% CI, −0.16% to 0.05% points), and rate of disagreements was 3.23% (9359 of 289 785; 95% CI, 3.17%-3.29%) in the intervention group and 3.28% (9533 of 290 670; 95% CI, 3.22%-3.35%) in the control group (difference, −0.05% points; 95% CI, −0.14% to 0.04% points).

Exploratory post hoc analysis showed that cancer detection rate for individual readers did not change with time spent on task, as represented by near identical odds of detecting cancer between the first and 40th case (OR, 0.987; 95% CI, 0.929-1.048). Results were very similar in the model adjusted for the characteristics of the woman screened (OR, 0.995; 95% CI, 0.938-1.055; eTable 5 in Supplement 2).

Exploratory post hoc analysis showed that recall rate for individual readers (the proportion of women that 1 reader determined should be recalled) reduced with time on task. The odds of recall decreased over the course of examining 40 cases (OR, 0.83; 95% CI, 0.81-0.85). The reduction was similar in the model adjusted for woman's age and previous attendance (OR, 0.89; 95% CI, 0.87-0.91; eTable 6 in Supplement 2). The mean change over the course of 40 cases was a reduction in recall rate from 6.4% (position 1) to 4.6% (position 40), with the trend continuing in longer batches (**Figure 3**).

Further exploratory post hoc analysis indicated that there was also no effect of the intervention when readers were blinded to one another's decision. For all 366 824 cases read from the 16 centers, the second reader was blinded to the first reader's decision results. In those centers, the cancer detection rate was 0.88% (1603 of 181 482; 95% CI, 0.84%-0.93%) in the intervention group and 0.87% (1611 of 185 342; 95% CI, 0.83%-0.91%) in the control group (difference, 0.01% points; 95% CI, −0.05% to 0.07% points). Similarly recall rate was 4.23% (7669 of 181 482; 95% CI, 4.13%-4.32%) in the intervention group and 4.23% (7847 of 185 342; 95% CI, 4.14%-4.33%) in the control group (difference, −0.01% points; 95% CI, −0.14% to 0.12% points).

Table 3. Factors Associated With Cancer Detection Rate Identified by Multilevel Logistic Regression Models, Unadjusted and Adjusted for Age and Previous Screening Attendance

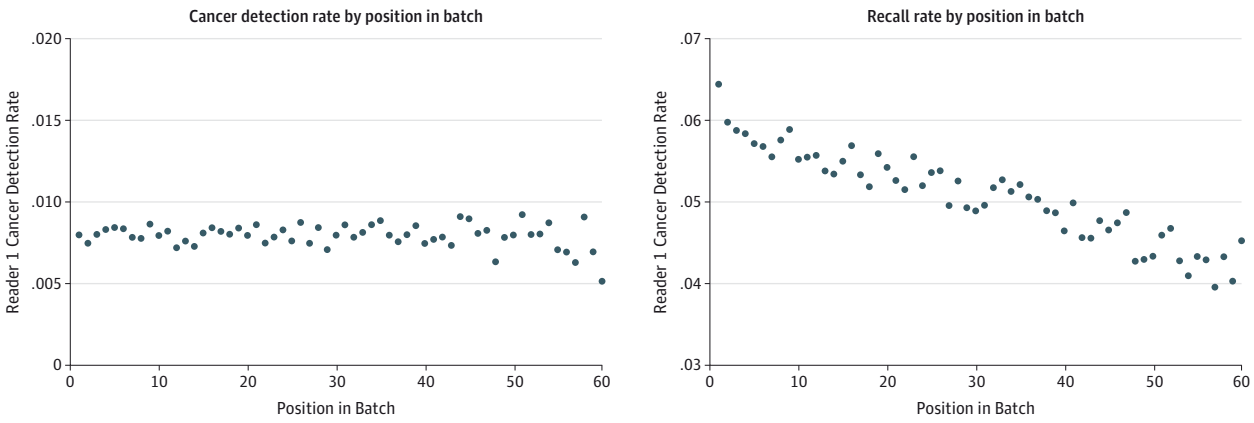
	Odd Ratio (95% CI)	
Variable	Unadjusted Model	Adjusted Model
Fixed-Effects (Measures of Association)		
Treatment variable		
Treatment vs control	1.01 (0.96-1.06)	1.01 (0.97-1.06)
Background factors		
Age, per year of age		1.052 (1.048-1.055)
No previous attendance		1.73 (1.62-1.86)
Random-Effects (Measures of Variation)		
Center level		
Variance, SE	0.058 (0.012-0.104)	0.038 (0.011-0.064)
Intracenter correlation, %	1.39	0.96
Median OR	1.26	1.20
P value for Wald statistics	.014	.006
Batch level		
Variance, SE	0.809 (0.754-0.863)	0.595 (0.543-0.647)
Intrabatch correlation, %	20.85	16.13
Median OR	2.35	2.08
P value for Wald statistics	<.001	<.001

Abbreviation: OR, odds ratio.

Discussion

This study examined whether an intervention to change the order in which readers examine breast screening cases could improve cancer detection rates. We randomized 1.2 million women in batches of approximately 35 to either intervention or control groups. The intervention did not influence cancer detection rate, recall rate, or rate of disagreement between readers. There was no pattern of decreasing cancer detection

Figure 3. Average Patterns of Cancer Detection Rate and Recall Rate for a Single Reader Over the Course of Examining a Batch of Mammograms



Each data point represents the mean recall or cancer detection rate over all cases examined by reader 1 at that position in the batch. A total of 1 173 930 cases were included, examined as reader 1 by 348 readers. The median number of screenings per batch position is 21 931 (interquartile range, 10 133-28 126).

rate with time on task as predicted by previous research on vigilance decrements as a psychological phenomenon. Instead there was a gradual decrease in recall rate, with an increase in PPV and a decrease in false-positive recall of women with time on task. This may reinforce and explain previous observational research that identifies that recall rate is reduced when grouping women's cases into batches.[23]

This randomized clinical trial was adequately powered to answer the research questions, with more than half of the English breast screening service participating. Effects were measured in a wide range of hospitals, increasing generalizability. Integration into the existing computer systems and reporting mechanisms resulted in very little loss to follow-up (<0.1%). Design of the trial computer system was iterative with high user involvement, which increased practicality and facilitated recruitment.

This study has several limitations. First, the main limitation is that reading conditions were not controlled, so although effectiveness in screening practice was measured, efficacy in ideal conditions was not evaluated. In this large pragmatic trial, we aimed to measure the effects of the intervention applied to current clinical practice in the United Kingdom, and we did not control for or measure working conditions, some of which may affect whether there is a vigilance decrement. Second, all readers would have met the minimum NHSBSP standards for reading volume, although we did not specify or measure the length of each reader's work week, the proportion of his/her time spent working in breast screening or reading mammograms, the number of work hours or type of work activities each day, number of breaks taken, or self-perceptions of fatigue. Similarly, although there are program-wide auditing methods for reader performance,[14] there will also be center-level variation in management of individual performance that we did not record. Third, the trial did not attempt to implement blinding of reader 2 to the decision of reader 1 in centers in which this was not standard practice, as limiting reader's access to computerized and paper notes was not considered possible without compromising patient safety. Fourth, 13% of women in the intervention group did not receive the intervention as intended. The trial software automatically detected these events, which occurred when readers manually overrode the case order and revisited the same case or used barcodes to identify individual cases. These women were included in the intention to treat analysis.

The trial results were unexpected and contradict previous research on the vigilance decrement in other fields.[5] The vigilance decrement phenomenon has been reported in many peer reviewed publications[5] but was not observed in this large randomized clinical trial. These previous studies were primarily undertaken in psychology laboratories rather than in real-life settings. Gur et al[24] demonstrated that performance in experimental conditions and in clinical practice may be very different, suggesting that there is a very different set of incentives in these 2 settings for the reader. Hancock[25] contends that the vigilance decrement is entirely a phenomenon created by the conditions designed to measure it. Another explanation for not observing any vigilance decrement is simply that the sessions were too short; however, batches of 40 cases take 20 to 30 minutes to examine,[26] and the vigilance decrement is usually complete 25 to 35 minutes into the task.[5] The experienced specialists in this study could be less prone to a vigilance decrement, as was found in experienced closed caption television operators reviewing a test film.[27] The vigilance decrement phenomenon may be associated with an increase in recall threshold rather than a reduction in performance.[28] If readers already have a low-recall threshold so that they are recalling cases with minimal indications of cancer on the mammograms, this may translate to an increase in specificity with minimal decrease in sensitivity. In addition, we have not yet tested the secondary outcome of the interval cancer rate (rate of cancers detected symptomatically between screening rounds). If there was a pattern in number of interval cancers with time on task, then this may provide evidence of a vigilance decrement. This will be investigated through future analysis of 3-year follow-up data. However, we are unlikely to observe such a pattern because the interval cancer rate is inversely proportional to the cancer detection rate and this does not change with time on task, and because all cases recalled by one reader received a reference standard of peers (independent examination by another reader followed by examination by a third reader or group of readers) and 60% received follow-up tests that included ultrasound and biopsy as appropriate. Furthermore, the increase in recall rate at the beginning of the batch is many times larger than the total number of interval cancers at screening.[1]

A reduction in recall rate with time spent on task has not previously been observed in breast cancer screening. However, an observational study has indicated that examining batches of women's mammograms in one sitting, rather than one by one reduces the overall recall rate with no change in cancer detection rate.[23]

The systematic reduction in recall rate with time on task for an individual reader did not translate into differences between the intervention and control group (double reading) in overall recall rate or rate of disagreements between readers. There are several possible explanations. The mechanism of action is dependent on the increased recall rates acting on the same cases in the control group and different cases in the intervention group. However, the situation is complex. Different readers have different recall thresholds and have different abilities to detect each type of mammographic abnormality (eg, spiculated masses, asymmetries, architectural distortions). Furthermore, each mammogram has overlapping tissue and many features that may appear suspicious. Therefore, for any particular pair, the increase in cases recalled at the beginning of the session may not manifest in recalling the same cases. If this is the case, then the intervention would not affect overall recall rate but it would affect who is recalled, with more women recalled at the beginning of the batch in the control group and with recalls spread more evenly throughout the batch in the intervention group.

The implications for practice are 2-fold. First the intervention of 2 readers examining a batch of mammograms in the opposite rather than the same order was not effective in increasing cancer detection rate. We have found no evidence

of harms from the intervention; however, some participating readers reported that it was more difficult to examine cases in reverse order because they also had to reverse associated paperwork. This result is generalizable only to population screening programs that use 2 readers to examine mammograms separately. These include the UK NHS breast screening programs for which double reading of mammograms was recommended and became mandatory following the transition to fully digital mammography,[14] European population screening program for which double reading is recommended and implemented,[29] and Australia where double reading is considered preferable[30] because it increases sensitivity[31] but is not mandated. In the United States, the Mammography Quality Standards Act and the US Food and Drug Administration do not require double reading of mammograms. The decision is made by professional societies and individual centers; in practice it rarely happens.

Second, for individual readers recall rate decreased with time spent on task for up to 60 cases, with no concurrent change in cancer detection rate. Therefore, we suggest that examining cases in batches of up to 60 is likely to be beneficial. This result was found across 360 readers, encompassing more than half of the NHS Breast Screening Programme in England. Therefore, it is likely to be generalizable to screening in England, and may be generalizable across all breast screening programs using batch reading. Examining mammograms in batches is now standard practice in high-volume population breast screening programs worldwide, with evidence that batch reading increases specificity.[23] However, batch reading is not always used, particularly when case volumes are low, such as in practices serving smaller populations. Batch reading is routine for other imaging studies not involving direct radiologist-patient contact with radiology information systems designed for this practice.

Conclusions

Interpretation of batches of mammograms by qualified screening mammography readers using a different order vs the same order for the second reading resulted in no significant difference in rates of detection of breast cancer.

ARTICLE INFORMATION

Author Contributions: Dr Taylor-Phillips had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Study concept and design: Taylor-Phillips, Wallis, Parsons, Dunn, Szczepura, Gates, Duncan, Hudson, Clarke.
Acquisition, analysis, or interpretation of data: Taylor-Phillips, Jenkinson, Adekanmbi, Parsons, Dunn, Stallard, Szczepura, Gates, Kearins, Duncan, Hudson, Clarke.
Drafting of the manuscript: Taylor-Phillips, Wallis, Jenkinson, Adekanmbi, Parsons, Clarke.
Critical revision of the manuscript for important intellectual content: Wallis, Jenkinson, Adekanmbi, Dunn, Stallard, Szczepura, Gates, Kearins, Duncan, Hudson, Clarke.
Statistical analysis: Taylor-Phillips, Jenkinson, Adekanmbi, Parsons, Stallard, Gates.
Obtained funding: Taylor-Phillips, Dunn, Szczepura, Clarke.
Administrative, technical, or material support: Taylor-Phillips, Wallis, Hudson, Clarke.
Study supervision: Taylor-Phillips, Wallis, Dunn, Szczepura, Duncan, Clarke.

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Dr Taylor-Phillips received postdoctoral fellowship funding from the UK National Institute of Heath Research (NIHR) to conduct the research and grant support from the UK National Screening Committee. Drs Wallis and Duncan work within the English NHS Breast Screening Programme. Dr Hudson's employers received payment for the time Dr Hudson spent developing the NBSS extracts for this research. Dr Kearins is the UK national lead for breast screening quality assurance and is employed by Public Health England. Drs Taylor-Phillips and Clarke currently receive funding for specified work on development of screening programs from Public Health England.

Funding/Support: This study presents independent research that was im part funded by the NIHR through a postdoctoral fellowship for Dr Taylor-Phillips. The research is also supported by the NIHR CLAHRC West Midlands initiative and sponsored by the University of Warwick.

Role of the Funder/Sponsor: The funders 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.

Disclaimer: The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health.

Additional Contributions: We thank David Solomon, BEc, and John Orrell, BSc, at Hitachi Consulting for adapting the NBSS software, Margot Wheaton at University Hospital Coventry and Warwickshire for her help designing the data extract, and the significant contributions of local collaborators and participants at the following breast screening centers: Avon (Bristol), Bedfordshire and Hertfordshire, Bolton, Bury, and Rochdale, Cambridge and Huntingdon, Canterbury, Chester City, Sandwell and Walsall, Cornwall, Derby City and South Derbyshire, Doncaster, Dudley and Wolverhampton, East Cheshire and Stockport, East Lancashire, East Suffolk, East Sussex Brighton and Hove, Gloucestershire, Guildford, Humberside, Kings Lynn, Leeds/Wakefield, Liverpool, Maidstone, Manchester, Medway, Mid Cheshire, Norfolk and Norwich, North Lancashire and South Cumbria, North Staffordshire, North Tees, North Yorkshire, Nottingham, Oxford, Portsmouth, Rotherham, Sheffield, Shropshire, Somerset, South Devon, South Staffordshire, South Lancashire, South West London, Southampton and Salisbury, Warrington, Warwickshire, Solihull and Coventry, West Devon and East Cornwall, and Wiltshire. The employer of Messrs Solomon and Orwell received financial reimbursement for their work developing the study software. None of the study centers nor Ms Wheaton received any financial reimbursement related to the study.

REFERENCES

1. Rayat P. Breast Screening Programme England statistics for 2014-15. London, England. http://www.hscic.gov.uk/article/2021/Website-Search?productid=20270&q=breast+screening&sort=Relevance&size=10&page=1&area=both#top. Published February 24, 2016. Accessed April 21, 2016.

2. National Institutes of Health. Breast Cancer Surveillance Consortium. Website. http://breastscreening.cancer.gov. Accessed February 10, 2015.

3. Bennett RL, Sellars SJ, Moss SM. Interval cancers in the NHS breast cancer screening programme in England, Wales, and Northern Ireland. *Br J Cancer*. 2011;104(4):571-577.

4. Mackworth N. The breakdown of vigilance during prolonged visual search. *Q J Exp Psychol*. 1948;1(1):6-21.

5. See JE, Howe SR, Warm JS, Dember WN. Meta-analysis of the sensitivity decrement in vigilance. *Psychol Bull*. 1995;117(2):230-249.

6. Badalamente RV, Ayoub MM. A behavioral analysis of an assembly line inspection task. *Human Factors*. 1969;11(4):339-352.

7. Basner M, Rubinstein J, Fomberstein KM, et al. Effects of night work, sleep loss and time on task on simulated threat detection performance. *Sleep*. 2008;31(9):1251-1259.

8. Verster JC, Roth T. Vigilance decrement during the on-the-road driving tests: the importance of time-on-task in psychopharmacological research. *Accid Anal Prev*. 2013;58:244-248.

9. Wiggins MW. Vigilance decrement during a simulated general aviation flight. *Appl Cogn Psychol*. 2011;25(2):229-235.

10. Gunn DV, Warm JS, Nelson WT, Bolia RS, Schumsky DA, Corcoran KJ. Target acquisition with UAVs: vigilance displays and advanced cuing interfaces. *Human Factors*. 2005;47(3):488-497.

11. Taylor-Phillips S, Elze MC, Krupinski EA, et al. Retrospective review of the drop in observer detection performance over time in lesion-enriched experimental studies. *J Digit Imaging*. 2015;28(1):32-40.

12. Gale A, De Silva E, Walker G, Roebuck E, Worthington B. Vigilance decrement and radiological reporting. Megaw ED, ed. *Contemporary Ergonomics*. Philadelphia, PA: Taylor & Francis; 1989:461-467.

13. Taylor-Phillips S, Wallis MG, Parsons H, et al. Changing case Order to Optimise patterns of Performance in mammography Screening (CO-OPS): study protocol for a randomized controlled trial. *Trials*. 2014;15:1-7.

14. 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. NHSBSP publication 59.

15. Baxter G, Jones V, Milnes V, et al *Guidance Notes for Equipment Evaluation and Protocol for User Evaluation of Imaging Equipment for Mammographic Screening and Assessment: NHSBSP Equipment Report 1411*. London, England: Cancer Screening Programmes, National Health Service; September 2014.

16. Health and Social Care Information Centre, Screening and Immunisations team. *Breast Screening Programme England Statistics for 2011-12*. Bristol, England: Centre for Multilevel Modelling, University of Bristol; February 2013.

17. Rasbash J, Steele F, Browne W, Prosser B. *A User's Guide to MLwiN*. Bristol, England: Center for Multilevel Modelling, University of Bristol; 2015.

18. Stata statistical software [computer program]. College Station, TX: StataCorp; 2015.

19. Goldstein H. *Multilevel Statistical Models*. London, England: Hodder Arnold; 2003.

20. Merlo J, Chaix B, Yang M, Lynch J, Råstam L. A brief conceptual tutorial of multilevel analysis in social epidemiology: linking the statistical concept of clustering to the idea of contextual phenomenon. *J Epidemiol Community Health*. 2005;59(6):443-449.

21. Snijders T, Bosker R. *Multilevel Analysis—An Introduction to Basic and Advanced Multilevel Modelling*. Thousand Oaks, California: SAGE publications; 1999.

22. Larsen K, Merlo J. Appropriate assessment of neighborhood effects on individual health: integrating random and fixed effects in multilevel logistic regression. *Am J Epidemiol*. 2005;161(1):81-88.

23. Burnside ES, Park JM, Fine JP, Sisney GA. The use of batch reading to improve the performance of screening mammography. *AJR Am J Roentgenol*. 2005;185(3):790-796.

24. Gur D, Bandos AI, Cohen CS, et al. The "laboratory" effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology*. 2008;249(1):47-53.

25. Hancock PA. In search of vigilance: the problem of iatrogenically created psychological phenomena. *Am Psychol*. 2013;68(2):97-109.

26. 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.

27. Donald F, Donald C, Thatcher A. Work exposure and vigilance decrements in closed circuit television surveillance. *Appl Ergon*. 2015;47:220-228.

28. Broadbent DE, Gregory M. Effects of noise and of signal rate upon vigilance analysed by means of decision theory. *Human Factors*. 1965;7(2):155-162.

29. 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. *Ann Oncol*. 2008;19(4):614-622.

30. Policy on mammography screening for breast cancer version 2. City, Country: The Royal Australian and New Zealand College of Radiologists; 2014.

31. Taylor P, Potts HWW. 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.