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[1]* PhD	s.taylor-phillips@warwick.ac.uk
Matthew G Wallis[2] FRCR	matthew.wallis@addenbrookes.nhs.uk
David Jenkinson[1] PhD	D.J.Jenkinson@warwick.ac.uk
Victor Adekanmbi[1] PhD	V.Adekanmbi@warwick.ac.uk
Helen Parsons[1] PhD	H.Parsons@warwick.ac.uk
Janet Dunn[1] PhD	J.A.Dunn@warwick.ac.uk
Nigel Stallard[1] PhD	N.Stallard@warwick.ac.uk
Ala Szczepura[1] DPhil	ala.szczepura@coventry.ac.uk
Simon Gates[1] PhD	S.Gates@warwick.ac.uk
Olive Kearins[3] MSc	Olive.Kearins@phe.gov.uk
Alison Duncan[4] FRCR	alison.duncan@uhcw.nhs.uk
Sue Hudson[5] MSc	sue.hudson@pasconsulting.co.uk
Aileen Clarke[1] MD	Aileen.Clarke@warwick.ac.uk

1. Warwick Medical School, the University of Warwick, Coventry, CV4 7AL, UK

2. Cambridge Breast Unit, Cambridge Universities NHS Foundation Trust and NIHR Cambridge Biomedical Research Centre, Cambridge CB2 0QQ, UK

3. Screening QA Service (Midland & East), Seaton House, City Link, Nottingham, NG2 4LA

4. Warwickshire, Solihull and Coventry Breast Screening Service, University Hospitals Coventry and Warwickshire, Coventry CV2 2DX

5. Peel & Schriek Consulting Limited, London N5 1AU

*Corresponding author tel (0044) 7725000262

Date of revision: 7th April 2016

Word count: 4846

Effect of using the same vs different order for second readings of screening mammograms on rates of breast cancer detection: A randomized clinical trial

Importance: Interpreting breast screening mammograms is a difficult repetitive task that can result in missed cancers and false positive recalls. In the UK, two 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 multi-centre, double-blind, cluster randomised controlled trial conducted at 46 specialised breast screening centres from the National Health Service Breast Screening Programme in England for 1 year, (all between 20th December 2012 and 3rd November 2014). 360 readers participated (mean 7.8 readers per centre), 186 radiologists, 143 radiography advanced practitioners and 31 breast clinicians, all fully qualified to report mammograms in the NHS breast screening programme.

Intervention: The two 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 (596642 in the intervention group; 597505 in the control group), the images were interpreted in 37,688 batches [median batch size 35 (IQR 16-46)], with each reader interpreting median 176 batches (IQR 96 to 278). After completion of all subsequent diagnostic tests, a total of 10,484 cases of breast cancer were detected (0.88%). There was no significant difference in cancer detection rate [5272 (0.88%) vs 5212 (0.87%), difference 0.011% points 95%CI -0.022 to 0.045], recall rate, [24681 (4.14%) vs 24894 (4.17%), difference -0.030% points 95%CI -0.101 to 0.042) or rate of reader disagreements [20,471 (3.43%) vs 20793 (3.48%), difference -0.048% points 95%CI -0.113 to 0.018,] between intervention and control groups.

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

Trial Registration: www.isrctn.com identifier: ISRCTN46603370

Introduction

Breast cancer screening currently detects 8.6 cancers per thousand women screened triennially (equivalent to 18,000 cancers per year) in the UK[1] and 4.2 cancers per thousand women screened annually in the US. [2] However, another 2.9 cancers per thousand women screened in the UK [3] (equivalent to 6,030 cancers per year) [1] and 0.9 cancers per thousand women screened in the US are detected between screening rounds in screened women. [2] These arise through cancers growing between screening rounds, and cancers missed at screening. An additional 3.4% of women in the UK (70,715 each year) [1] and 9.3% of women in the US [2] experience false positive recalls at each screening round.

Interpreting screening mammograms is a difficult and repetitive visual search task, where characteristics of cancer are disguised amongst 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 aeroplanes[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 UK two film readers independently examine each woman's mammograms for signs of cancer. In this study we investigated whether there is a vigilance decrement to detect cancer in breast screening practice, and whether changing the order in which the two 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) pragmatic double blind cluster randomised controlled trial was designed to determine whether there is a vigilance decrement in breast cancer screening, and whether changing the order in which the cases are presented can increase cancer detection rate.

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

Intervention and Outcomes

The study compared two parallel groups, each split into two sub-groups to ensure blinding of the readers. The intervention group involved the two readers reading the batch in the opposite order to each other; one forwards, one in reverse. Hence, the two sub-groups: first reader forwards, 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 sub-groups being either both forwards (which is current practice) or both in reverse (to maintain the blinding of a reader to trial group, as they would be aware that they are reading a batch in reverse). Thus each batch (cluster) was randomised with equal probability to one of four groups.

The primary outcome was cancer detection rate, (number of women with cancer detected as a proportion of all women screened) as 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 one reader results in high vigilance states occurring for the two readers when examining different women's mammograms, so the cancers are detected by at least one of the experts, as outlined in figure 1. If a reader in a high vigilance state detected a cancer missed by their 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

Centres 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 centres 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 centres in England which took part in the trial in comparison to those which did not is provided in table e1. The trial ran for 1 year at each centre, with individual centres starting the study when local consent and research and development approvals were obtained, (start dates were all between 20th December 2012 and 4th November 2013). One centre completed only 4 months of the study due to local technical and workforce issues.

Mammograms from women attending routine breast cancer screening at these centres during the study period were included. These were arranged into batches of around 40 women as is standard practice in the UK, and 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 centre level, with consent of individual women considered impractical for this system level intervention. In the UK women age 50 to 70 are invited to breast screening every 3 years, this study also includes women aged 47-49 and 71-73 participating in the age extension trial (NCT01081288), and a small proportion of older women (2.3% of women in the trial) who self-refer as part of the programme. 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 5,000 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 (PERFORMS) test set.[14] Each centre 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 two readers co-located in the same breast screening centre. 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 centres workflow systems were designed to blind reader 2 to the decision of reader 1. All centres used arbitration when the two readers disagreed, with 13 centres using a single 3rd reader, and 33 centres using group consensus of 2 or more readers.

Randomisation and Blinding

The randomisation 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 randomisation 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 randomisation was a batch of mammograms, whereas the unit of observation was the individual mammogram. Simple randomisation was used without stratification or minimisation due to the large number of clusters randomised.

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 centre's annual reporting requirements, to reduce missing data. The data was extracted through NBSS from each centre, exporting data in Excel format. The datasets from each centre were merged using Excel and R [v 3.0.3 in RStudio v 0.98.501]. Cancer was defined as needle biopsy or surgery 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 2011-12) the breast cancer detection rate in the UK 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 eight breast screening centres in one English region. This suggested that the intervention may result in one 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 randomisation being the batch, so the sample size needed to be inflated by the design effect. The inter-cluster 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 centres. There were no interim analyses or stopping rules.

Statistical analysis

We used multivariable multilevel logistic regression to analyse factors associated with breast cancer detection, recall and disagreement rates due to the hierarchical nature of the datasets. 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 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) while the fourth model added the intervention to the adjusted model. All multilevel modelling was performed using MlwiN 2.35 [17] called from Stata statistical software for Windows version 14 [18] using runmlwin routine. For the multilevel logistic regression models, (iterative generalized least squares; penalised quasi-likelihood) IGLS PQL2 estimation was used. [19] Two-tailed tests were used, with p values <0.05 considered significant. The fixed effects (i.e. measures of association) are presented as adjusted odds ratios with their corresponding 95% confidence intervals (CIs). Measures of random effects included intra-cluster correlation (ICC) and median odds ratio (MOR). [20, Merlo, #9] The ICC was calculated by the linear threshold according to the formula used by Snijders et. al. [21] while MOR is a measure of unexplained cluster heterogeneity. Methods used for calculating MOR have been described elsewhere [20, Merlo, #9,22]. 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 three pre-defined sub-groups: women aged under 53 (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 (where 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 sub group 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 (i.e. 1st, 2nd, 3rd …) 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 personalised 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 modelling 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 analysed. 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 centres. Including only the sub-group of centres 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 to those in the control group.

Results

Flow of Women in the CO-OPS Trial

1,207,633 women were included in the trial, see figure 2. There were three causes of loss to follow up: 258 (0.02%) were recalled for further tests from screening but did not attend, 233 (0.02%) had an inconclusive needle biopsy result but refused further tests, and 298 women (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 moved house and GP practice and consequently were re-invited 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 (quartiles 16 and 46). Each reader examined a median of 5640 cases, (IQR 2599 to 8458), in a median of 176 batches (IQR 96 to 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 to 4). Each centre examined between 8152 and 72714 cases (median 25540 cases).

Outcomes

The primary outcome, cancer detection rate, was 0.88% (5272/596642) in the intervention group and 0.87% (5212/597505) in the control group (difference 0.011% points 95%CI -0.022 to 0.045), see table 2. The intervention did not affect cancer detection rate in the unadjusted (OR=1.01 95%CI 0.96 to 1.06) or adjusted models (OR=1.01 95%CI 0.97 to 1.06), see table 3 and e2. In the adjusted model, cancer detection rate increased with each increasing year of age (OR=1.052 95%CI 1.048 to 1.055) and was higher in women who had not previously attended screening (OR=1.73 95%CI 1.62 to 1.86).

The intervention also had no effect in any of the sub-groups of younger age, first and last 5 cases in the 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% (580/70071, 95%CI 0.76% to 0.89%) in the intervention group and 0.88% (623/70715, 95%CI 0.81% to 0.95%) in the control group (difference -0.053% points 95%CI -0.149 to 0.043). For batches read second or subsequent in each workday by both readers cancer detection rate was 0.85% (2472/289786, 95%CI 0.82% to 0.89%) in the intervention group and 0.85% (2473/290671, 95%CI 0.82% to 0.88%) in the control group (difference 0.002% points 95%CI -0.045 to 0.050).

The intervention did not affect either of the secondary outcomes, recall rate or rate of disagreements. The recall rate was 4.14% (24681/596642) in the intervention group and 4.17% (24894/597505) in the control group (difference -0.030% points 95%CI -0.101 to 0.042), see table 2. The rate of disagreement was 3.43% in the intervention group (20471/596294) and 3.48% (20793/597387) in the control group (difference -0.048% points 95%CI -0.113 to 0.018), as detailed in table 2. The intervention had no effect on recall rate in the unadjusted (OR=0.993 CI 0.974 to 1.013) or adjusted (OR=0.997 CI 0.978 to 1.016) models, (see supplementary table e3) or on rate of disagreement in the unadjusted (OR=0.994 CI 0.971 to 1.019) or adjusted model (OR=0.997 CI 0.974 to 1.020), see supplementary table e4. Recall rate was higher with each year of age of the woman screened (OR 1.008 CI 1.007 to 1.010), and was higher in women who had not previously attended breast screening (OR=2.89 CI 2.82 to 2.97). Rate of disagreement was also higher for women at their first screening appointment (OR=2.17 CI 2.11 to 2.24) but lower with each year of increasing age of the woman screened (OR=0.994 CI 0.992 to 0.996). The positive predictive value (PPV) was 21.4% (95% CI 20.8% to 21.9%) in the intervention group and 20.9% (95% CI 20.4% to 21.4%) in the control group (difference 0.420% points 95%CI -0.299 to 1.139). The intervention had no effect on any of the sub-groups (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 recall rate was 4.02% (2818/70071, 95%CI 3.88% to 4.17%) in the intervention group and 4.11% (2904/70715, 95%CI 3.96% to 4.25%) in the control group (difference -0.085% points 95%CI -0.291 to 0.121), and rate of disagreements was 3.61% (2531/70071, 95%CI 3.47% to 3.75%) in the intervention group and 3.75% (2653/70715, 95%CI 3.61% to 3.89%) in the control group (difference -0.140% points 95%CI -0.336 to 0.057). For batches read second or subsequent in each workday by both readers recall rate was 4.10% (11868/289786, 95%CI 4.02% to 4.17%) in the intervention group and 4.15% (12068/290671, 95%CI 4.08% to 4.22%) in the control group (difference -0.056% points 95%CI -0.159 to 0.046), and rate of disagreements was 3.23% (9359/289785, 95%CI 3.17% to 3.29%) in the intervention group and 3.28% (9533/290670, 95%CI 3.22% to 3.35%) in the control group (difference -0.050% points 95%CI -0.141 to 0.041).

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 fortieth case (OR=0.987, 95%CI 0.929 to 1.048). Results were very similar in the model adjusted for the characteristics of the woman screened (OR=0.995 95%CI 0.938 to 1.055), supplementary table e5.

Exploratory post-hoc analysis showed that recall rate for individual readers (the proportion of women that one 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 to 0.85). The reduction was similar in the model adjusted for woman's age and previous attendance (OR=0.89 95%CI 0.87 to 0.91), see supplementary table e6. 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, see 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. 366,824 cases were read in the trial at the 16 centres which blind reader 2 to reader 1 decision. In those centres the cancer detection rate was 0.88% (1603/181482, 95%CI 0.84% to 0.93%) in the intervention group and 0.87% (1611/185342, 95%CI 0.83% to 0.91%) in the control group (difference 0.014% points, 95% CI -0.046 to 0.074). Similarly recall rate was 4.23% (7669/181482 95%CI 4.13% to 4.32%) in the intervention group and 4.23% (7847/185342, 95%CI 4.14% to 4.33%) in the control group (difference -0.008% points, 95%CI -0.138 to 0.122).

Discussion

We examined whether an intervention to change the order in which readers examine breast screening cases could improve cancer detection rate. We randomised 1.2 million women in batches of approximately 35 to 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 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 positive predictive value and a decrease in false positive recall of women with time on task. This may reinforce and explain previous observational research which identifies that recall rate is reduced when grouping women's cases into batches. [23]

This randomised controlled trial in 1.2 million women was adequately powered to answer the research questions, with over half of the English breast screening service taking part. Effects were measured in a wide range of hospitals, increasing generalisability. Integration into the existing computer systems and reporting mechanisms resulted in very little loss to follow-up (less than 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 whilst 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 UK, 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 readers work week, the proportion of their 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 whilst there are programme wide auditing methods for reader performance,[14] there will also be centre level variation in management of individual performance which we did not record. Third, the trial did not attempt to implement blinding of reader 2 to the decision of reader 1 where this was not standard practice, as limiting reader's access to computerised 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 randomised controlled 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 two settings for the reader. Hancock contends that the vigilance decrement is entirely a phenomenon created by the conditions designed to measure it. [25] Another explanation for not observing any vigilance decrement is simply that the sessions were too short, however, batches of 40 cases take 20-30 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 CCTV 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] and if readers already have a low recall threshold so 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 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 interval cancer rate is inversely proportional to cancer detection rate and this does not change with time on task, and because all cases recalled by one reader did receive 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 which 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 upon the increased recall rates acting upon 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 different abilities to detect each type of mammographic abnormality (eg spiculated masses, asymmetries, architectural distortions etc). Furthermore each mammogram has overlapping tissue and many features which 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 recalls spread more evenly throughout the batch in the intervention group.

The implications for practice are two-fold. Firstly the intervention of two readers examining a batch of mammograms in the opposite rather than the same order is 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 as they also had to reverse associated paperwork. This result is only generalizable to population screening programmes which use two readers to examine mammograms separately. These include the UK NHS breast screening programmes where double reading of mammograms was recommended and became mandatory following the transition to fully digital mammography,[14] European population screening programmes where double reading is recommended and implemented, [29] and Australia where double reading is considered preferable[30] because it increases sensitivity,[31] but not mandated. In the US the Mammography Quality Standards Act and the FDA do not require double reading of mammograms, the decision is made by professional societies and individual centres, and in practice it rarely happens.

Secondly, for individual readers recall rate decreases 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 generalisable to screening in England, and may be generalisable across all breast screening programmes using batch reading. Examining mammograms in batches is now standard practice in high volume population breast screening programmes 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.

Conclusion

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

Acknowledgements

This paper presents independent research funded by the National Institute for Health Research (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. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript

We are very grateful for the contribution of David Solomon BEc and John Orrell BSc at Hitachi Consulting for adapting the NBSS software, to Margot Wheaton at University Hospital Coventry and Warwickshire for her help designing the data extract, and to the significant contributions of local collaborators and participants at the following breast screening centres: Avon (Bristol), Bedfordshire & Hertfordshire, Bolton, Bury & Rochdale, Cambridge & Huntingdon, Canterbury, Chester, City, Sandwell & Walsall, Cornwall, Derby City & South Derbyshire, Doncaster, Dudley & Wolverhampton, East Cheshire & Stockport, East Lancashire, East Suffolk, East Sussex Brighton and Hove, Gloucestershire, Guildford, Humberside, Kings Lynn, Leeds/Wakefield, Liverpool, Maidstone, Manchester, Medway, Mid Cheshire, Norfolk & Norwich, North Lancashire & 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 & Salisbury, Warrington, Warwickshire, Solihull & Coventry, West Devon & East Cornwall, and Wiltshire. The employer of DS and JO received financial reimbursement for their work developing the study software, none of the study centres or MW received any financial reimbursement related to the study. Sian Taylor-Phillips had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Contributions

All authors contributed to the design of the study, and the write up. STP led the study including obtaining funding, design, data collection, analysis and write up. HP performed the sample size calculations. STP, SH and DJ collected the data. STP, DJ, VA and NS contributed to analysing the data.

Conflicts of Interest

STP received postdoctoral fellowship funding from the UK National Institute of Heath Research to conduct the research. MW and AD work within the English NHS Breast Screening Programme. SH's employers received payment for the time SH spent developing the NBSS extracts for this research. OK is the UK national lead for breast screening quality assurance, employed by Public Health England. STP and AC currently receive funding for specified work on development of screening programmes from Public Health England.

References

1. Wilson R. *Breast Screening Programme England Statistics for 2013-14.* Screening and Immunisations team, Health and Social Care Information Centre; 18th February 2015. ISBN 978-1-78386-319-8.
2. NCI-funded Breast Cancer Surveillance Consortium (HHSN261201100031C). Downloaded on 2/10/15 from the Breast Cancer Surveillance Consortium website http://breastscreening.cancer.gov/.
3. Bennett RL, Sellars SJ, Moss SM. Interval cancers in the NHS breast cancer screening programme in England, Wales and Northern Ireland. *British Journal of Cancer.* 2011;104(4):571-577.
4. Mackworth N. The breakdown of vigilance durning prolonged visual search. *Quarterly Journal of Experimental Psychology.* 1948;1(1):6-21.
5. See JE, Howe SR, Warm JS, Dember WN. Meta-analysis of the sensitivity decrement in vigilance. *Psychological Bulletin.* 1995;117(2):230.
6. Badalamente RV, Ayoub MM. A Behavioral Analysis of an Assembly Line Inspection Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society.* 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. *Accident Analysis & Prevention.* 2013;58:244-248.
9. Wiggins MW. Vigilance decrement during a simulated general aviation flight. *Applied Cognitive Psychology.* 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: The Journal of the Human Factors and Ergonomics Society.* 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. *Journal of digital 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) Contemporary Ergonomics.* 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;15:17.
14. Wilson R, Liston J. *Quality Assurance Guidelines for Breast Cancer Screening Radiology: NHS Breast Screening Programme Publication Number 59.* Sheffield: NHS Cancer Screening Programmes;2011.
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 version 4.* UK: Public Health England; September 2014 2014.
16. McConnell C. *Breast Screening Programme England Statistics for 2011-12.* Screening and Immunisations team, Health and Social Care Information Centre; 27 February 2013. ISBN 978-1-84636-837-0.
17. Rasbash J, Steele, F, Browne W, Prosser B. *A user's guide to MLwiN.* 2015.
18. *Stata statistical software* [computer program]. College Station, TX: StataCorp.; 2015.
19. Goldstein H. *Multilevel statistical models.* London: 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. *Journal of epidemiology and 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. *American journal of epidemiology.* 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. *American Journal of Roentgenology.* 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 P. In search of vigilance: the problem of iatrogenically created psychological phenomena. *The American Psychologist.* 2013;68(2):97-109.

26. Taylor-Phillips S, Wallis MG, Gale AG. Should previous mammograms be digitised in the transition to digital mammography? *European radiology.* 2009;19(8):1890-1896.

27. Donald F, Donald C, Thatcher A. Work exposure and vigilance decrements in closed circuit television surveillance. *Applied Ergonomics.* 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: The Journal of the Human Factors and Ergonomics Society.* 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. *Annals of Oncology.* 2008;19(4):614-622.

30. *Policy on Mammography Screening for Breast Cancer Version 2.* 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. *European Journal of Cancer.* 2008;44(6):798-807.

Table 1 Baseline characteristics for intervention and control groups

	Intervention	Control
Individual level		
Mean age of women screened (sd)	59.3 (7.48)	59.3 (7.49)
Number who have previously attended screening (%)	126,490 / 596,642 (21.2%)	128,217 / 597,505 (21.5%)
Cluster level		
Median batch length (quartiles)	35 (16,46)	35 (16,45)
Median number of screenings examined by each reader (quartiles)	2,848 (1,469 , 4,385)	2,891 (1,543 , 4,458)
Median number of batches examined by each reader (quartiles)	86 (52,143)	91 (51,138)
Median number of screenings examined at each centre (quartiles)	12,496 (8,997 , 16,523)	12,908 (9,529 , 16,418)
Median number of batches examined at each centre (quartiles)	376 (282,502)	364 (272,521)

Table 2. Primary and secondary outcomes in intervention and control groups.

Cancer detection rate, recall rate and rate of disagreement between readers in screenings of previous attenders, screenings of previous non-attenders and all screenings.

Outcome	**Intervention**	**Control**	**Difference**
Primary outcome: Cancer detection rate			
All screenings (CI)	0.88% (0.86% - 0.91%) 5,272/596,642	0.87% (0.85% - 0.90%) 5,212/597,505	0.011% points (-0.022 - +0.045)
Screenings of previous attenders (CI)	0.90% (0.87% - 0.92%) 4,214/470,152	0.88% (0.85% - 0.91%) 4,122/469,288	0.018% points (-0.020 - +0.056)
Screenings of previous non-attenders (CI)	0.84% (0.79% - 0.89%) 1,058/126,490	0.85% (0.80% - 0.90%) 1,090/128,217	-0.014% points (-0.085 - +0.057)
Secondary outcome: Recall Rate			
All screenings (CI)	4.14% (4.09% - 4.19%) 24,681/596,642	4.17% (4.12% - 4.22%) 24,894/597,505	-0.030% points (-0.101 - +0.042)
Screenings of previous attenders (CI)	3.15% (3.10% - 3.20%) 14,819/470,152	3.17% (3.12% - 3.22%) 14,869/469,288	-0.016% points (-0.087 - +0.054)
Screenings of previous non-attenders (CI)	7.80% (7.65% - 7.94%) 9,862/126,490	7.82% (7.67% - 7.97%) 10,025/128,217	-0.022% points (-0.231 - +0.186)
Secondary outcome: Disagreement rate between readers			
All screenings (CI)	3.43% (3.39% - 3.48%) 20,471/596,294	3.48% (3.43% - 3.53%) 20,793/597,387	-0.048% points (-0.113 - +0.018)
Screenings of previous attenders (CI)	2.73% (2.69% - 2.78%) 12,850/469,869	2.76% (2.71% - 2.80%) 12,937/469,215	-0.022% points (-0.088 - +0.044)
Screenings of previous non-attenders (CI)	6.03% (5.90% - 6.16%) 7,621/126,425	6.13% (6.00% - 6.26%) 7,856/128,172	-0.101% points (-0.287 - +0.084)

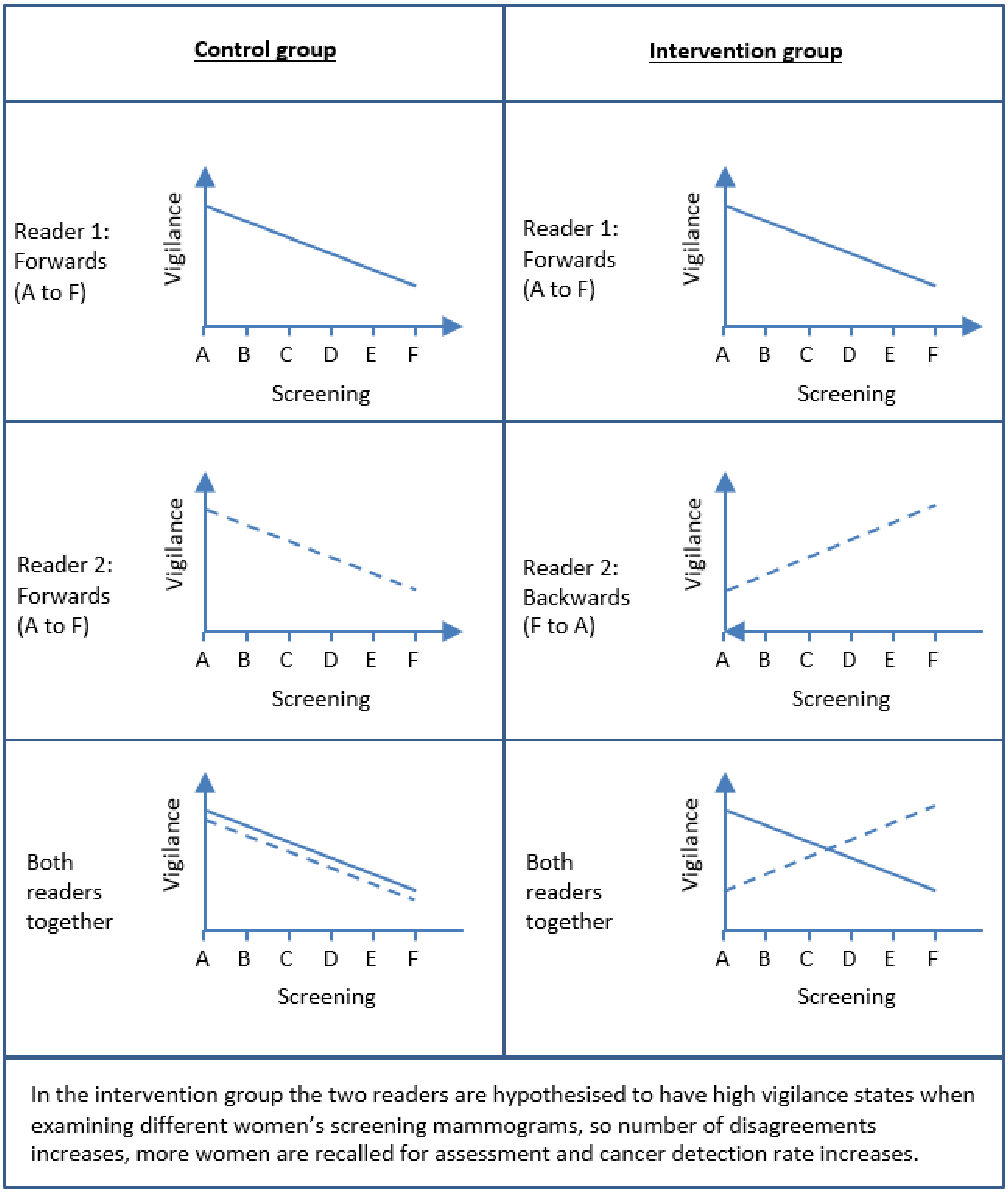
Table 3: Factors associated with cancer detection rate identified by multilevel logistic regression models, unadjusted and adjusted for age and previous attendance.

Variable	Unadjusted Model **OR (CI)**	Adjusted model **OR (CI)**
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)		
Centre level		
Variance (SE)	0.058(0.012-0.104)	0.038(0.011-0.064)
Intra-centre correlation (%)	1.39	0.96
MOR	1.26	1.20
Wald statistics (p-value)	0.014	0.006
Batch level		
Variance (SE)	0.809(0.754-0.863)	0.595(0.543-0.647)
Intra-batch correlation (%)	20.85	16.13
MOR	2.35	2.08
Wald statistics (p-value)	<0.001	<0.001

Abbreviations: SE; standard error, CI; confidence interval, MOR; median odds ratio

Figure Legends

Figure 1: Proposed mechanism of action of changing case order intervention, assuming the hypothesised vigilance decrement. Each screening represents examining a set of four mammograms, mediolateral oblique and craniocaudal views of both breasts for one woman.



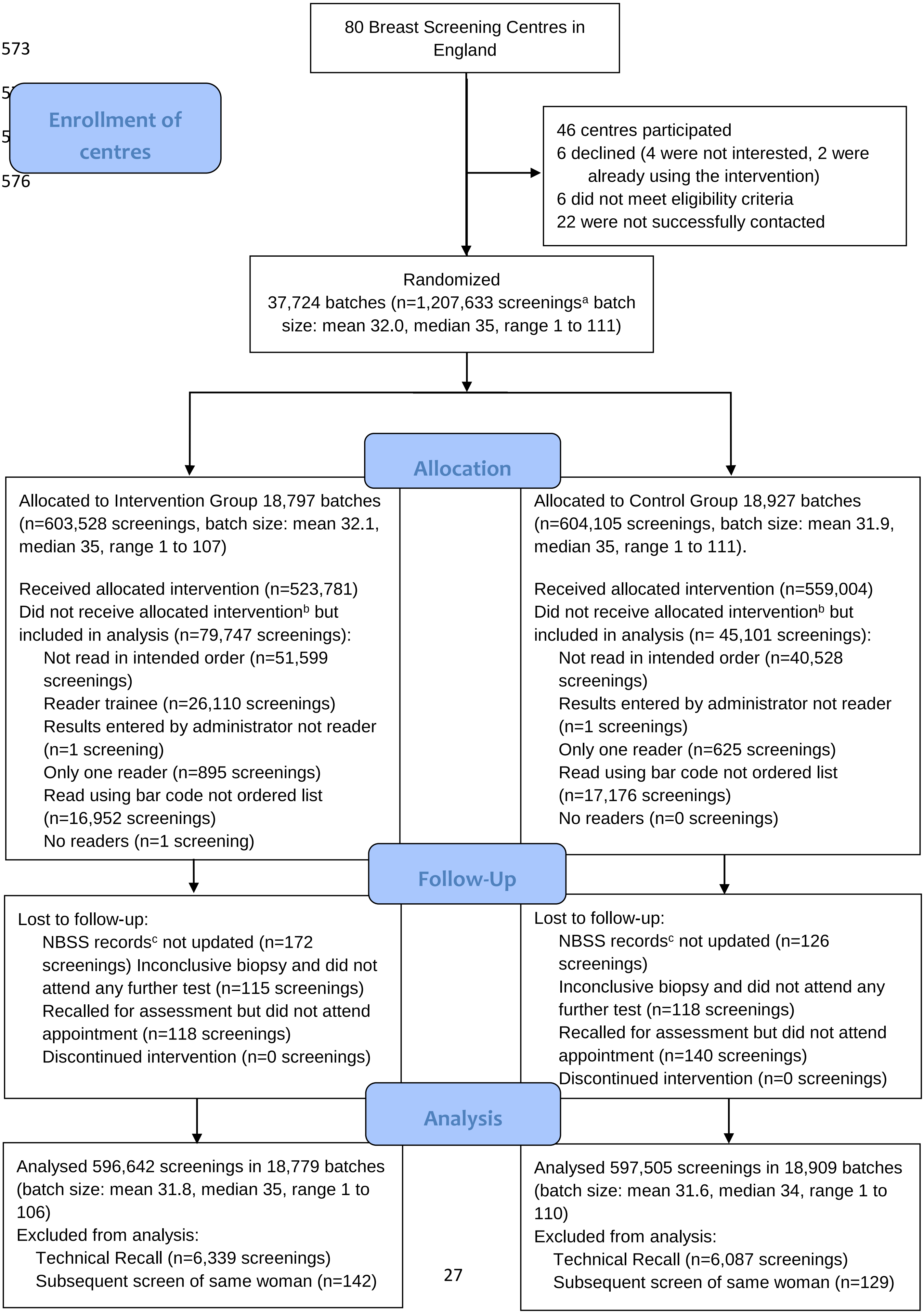
In the intervention group the two readers are hypothesised to have high vigilance states when examining different women's screening mammograms, so number of disagreements increases, more women are recalled for assessment and cancer detection rate increases.

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.



80 Breast Screening Centres in England

46 centres participated
6 declined (4 were not interested, 2 were already using the intervention)
6 did not meet eligibility criteria
22 were not successfully contacted

Randomized
37,724 batches (n=1,207,633 screenings[a] batch size: mean 32.0, median 35, range 1 to 111)

Allocated to Intervention Group 18,797 batches (n=603,528 screenings, batch size: mean 32.1, median 35, range 1 to 107)

Received allocated intervention (n=523,781)
Did not receive allocated intervention[b] but included in analysis (n=79,747 screenings):
Not read in intended order (n=51,599 screenings)
Reader trainee (n=26,110 screenings)
Results entered by administrator not reader (n=1 screening)
Only one reader (n=895 screenings)
Read using bar code not ordered list (n=16,952 screenings)
No readers (n=1 screening)

Allocated to Control Group 18,927 batches (n=604,105 screenings, batch size: mean 31.9, median 35, range 1 to 111).

Received allocated intervention (n=559,004)
Did not receive allocated intervention[b] but included in analysis (n= 45,101 screenings):
Not read in intended order (n=40,528 screenings)
Results entered by administrator not reader (n=1 screenings)
Only one reader (n=625 screenings)
Read using bar code not ordered list (n=17,176 screenings)
No readers (n=0 screenings)

Lost to follow-up:
NBSS records[c] not updated (n=172 screenings) Inconclusive biopsy and did not attend any further test (n=115 screenings)
Recalled for assessment but did not attend appointment (n=118 screenings)
Discontinued intervention (n=0 screenings)

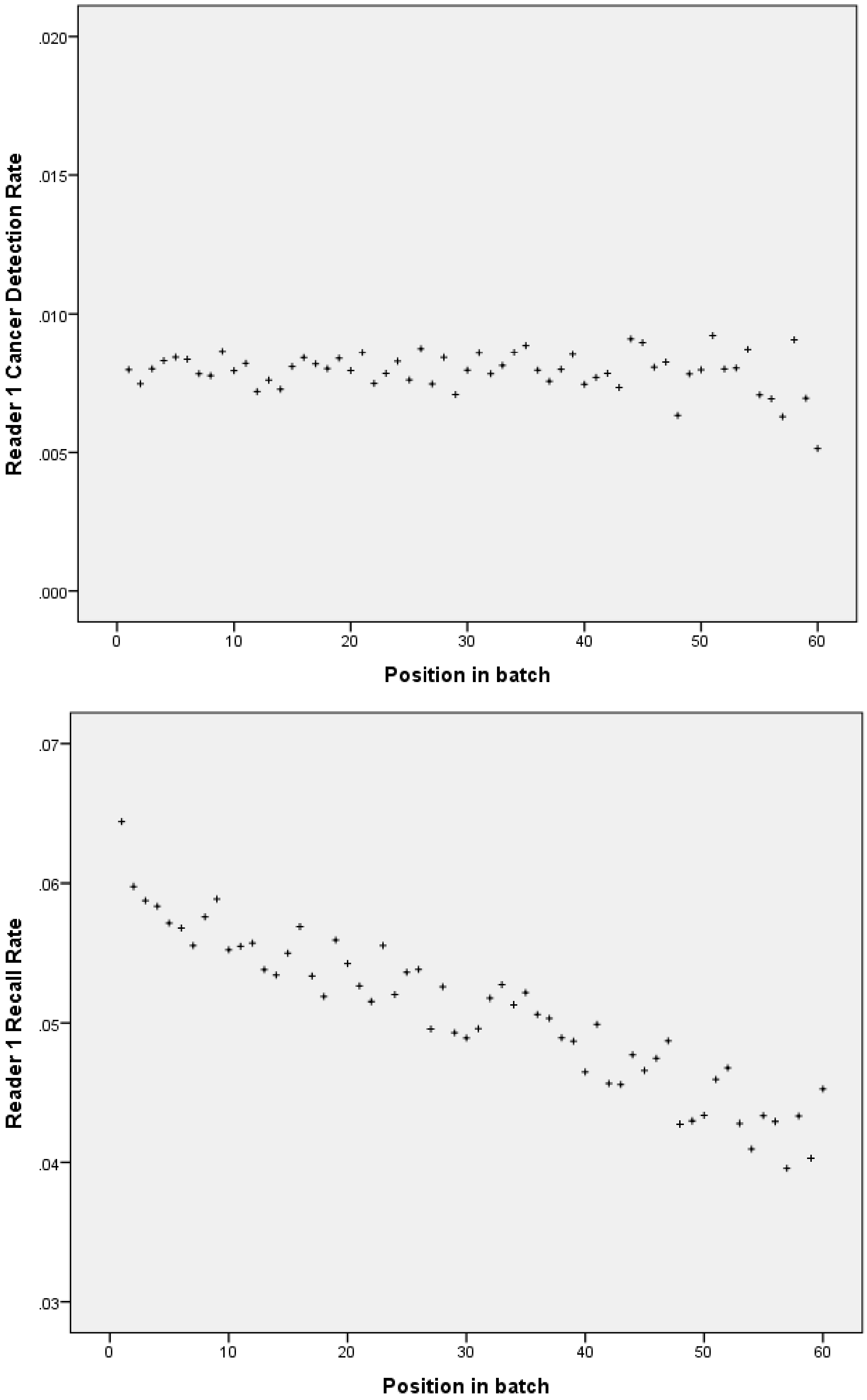
Lost to follow-up:
NBSS records[c] not updated (n=126 screenings)
Inconclusive biopsy and did not attend any further test (n=118 screenings)
Recalled for assessment but did not attend appointment (n=140 screenings)
Discontinued intervention (n=0 screenings)

Analysed 596,642 screenings in 18,779 batches (batch size: mean 31.8, median 35, range 1 to 106)
Excluded from analysis:
Technical Recall (n=6,339 screenings)
Subsequent screen of same woman (n=142)

Analysed 597,505 screenings in 18,909 batches (batch size: mean 31.6, median 34, range 1 to 110)
Excluded from analysis:
Technical Recall (n=6,087 screenings)
Subsequent screen of same woman (n=129)

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. 1,173,930 cases are included, examined as reader 1 by 348 readers, median number of screenings per batch position is 21,931 (IQR 10,133 to 28,126).


Reader 1 Cancer Detection Rate
Position in batch
Reader 1 Recall Rate
Position in batch