THE UNIVERSITY OF WARWICK

Warwick Medical School

Changing case Order to Optimise patterns of Performance in Screening (CO-OPS) Trial

RESEARCH PROTOCOL

Version 2.0
24th July 2013

NIHR Postdoctoral research fellowship for Dr Sian Taylor-Phillips reference NIHR-PDF-2011-01-073 Improving breast cancer screening detection rates through understanding, modelling, and adapting patterns of radiologist performance

Approved by the Coventry and Warwickshire NHS Research Ethics Committee on 27th June 2012 Reference 12/WM/0182

Sponsor: University of Warwick

1. Background

1.1 Epidemiology and burden of condition

Breast cancer is a leading cause of mortality in women, and was the cause of death for 10,280 women in England and Wales in 2010. It was responsible for 12% of all deaths of women in their fifties, and 8% of deaths of women in their sixties. [1]

In England 2.3 million women are projected to attend breast cancer screening each year, after the age extension of eligibility is increased to 47-73.[2,3] The cancer detection rate is currently 8 per thousand women screened. There are also 36 healthy women recalled for further tests (false positive recalls) per thousand screened.[2] 66% of all breast cancers in women aged 50-64 in England are detected by screening,[3] with the remaining 34% including women who choose not to attend screening, fast growing cancers which develop between screening rounds, and cancers missed at screening. The NHS Breast Screening programme (NHSBSP) aims to keep the number of cancers detected between screening rounds (interval cancers) down to 2.3 per thousand women,[4] however these are difficult targets and a more realistic rate may be more than 3.7 per thousand women.[4]

A related issue is that of over-diagnosis as a result of the introduction of breast cancer screening programmes, whereby cancers are detected at screening and treated, but this treatment is unnecessary because the woman would have died of other causes before the cancer ever grew to a stage where it would affect her health. It is very difficult to determine in the screening programme which cancers have been detected early and therefore saved the woman's life, and which cancers would never have threatened her life.

Therefore missed cancers (false negatives) and recall of healthy women for further tests (false positive recalls) are significant issues in mammography screening, but must be considered in the context of possible over-diagnosis.

1.2 Existing knowledge

In the NHSBSP, each set of mammograms is read by two radiologists who each indicate if there should be a recall for further tests. Disagreements are referred for arbitration, usually by a third radiologist. Radiologists assess two mammograms per minute on average[5] which equates to 120 in a 1-hour session. This is a highly skilled, pressurised but repetitive and frequently intellectually unchallenging activity. A vigilance decrement of performance decrease over time has been observed in similar repetitive visual tasks such as radar operator.[6]

Data from an observational study using the Breast Screening Programme database for eight English breast screening centres for 3 years has shown that the probability of being recalled for further tests and of having cancer detected at screening is dependent on the time since the radiologist had a break from reading. For the first ten cases after a first reader break, the cancer detection rate for the whole process (including second reader and arbitration) is higher. These extra cancers detected are not associated with a higher probability of being DCIS rather than invasive, and have a similar mix of grades 1, 2 and 3 to the rest of the cancers detected at screening so this phenomenon is unlikely to be associated with increased overdiagnosis.

First and second readers read each batch of mammograms in the same order, thus their post-break increases in cancer detection rate occur when reading the same women's mammograms.

1.3 Hypothesis

If the second radiologist examines the batch of mammograms in the reverse order to the first radiologist this will increase the overall cancer detection rates, because each radiologist's peak performance will be when examining different cases.

1.4 Need for a Trial

A previous observational study has shown a reduction in sensitivity with a proxy for time (number of cases) since a break. Currently both radiologists are presented with cases in the same order as one another, so both have reduced sensitivity when reading the same cases. Therefore changing the order in which the second radiologist is presented with the mammograms may increase sensitivity of screening. A trial is required to determine whether such improvements would be realised in practice.

1.5 Good Clinical Practice

The trial will be carried out in accordance with Good Clinical Practice (GCP) as detailed in the Medical Research Council (MRC) GCP guidelines
(http://www.mrc.ac.uk/Utilities/Documentrecord/index.htm?d=MRC002416)
and in accordance with the following protocol.

1.6 CONSORT

The trial will be reported in line with the CONSORT (*Consolidated Standards of Reporting Trials)* statement, and the further guidance available for cluster randomised controlled trials.

2 Trial Design

2.1 Trial summary

The trial is a multi-centre cluster randomised parallel group superiority trial. The centres will be English Breast Screening centres with digital mammography equipment. The intervention is reversing the order in which the second radiologist reads each batch of mammograms. A cluster for randomisation is a batch of mammograms (circa 30 women's mammograms). The unit for analysis is the individual women's outcome.

The pre-pilot will be conducted in three screening centres for two weeks to ascertain the practical issues with implementation, and to test the Crystal report for data extraction. Concurrently a survey of all English breast screening centres will be conducted to ascertain key characteristics relevant to the trial which may differ between centres, such as the blinding of one reader to another's decision, and the method of arbitration used.

The trial will be implemented using an adaptation to the NBSS software used at participating centres.

The trial will last 16 months in 44 centres (1.5 million women). The first 12 months data (1.1 million women) will be collected from the NBSS database to determine whether the intervention affected cancer detection rate.

The full data set will be collected three years after the end of the study, to ascertain whether there was a decrease in rate of interval cancers (cancers presenting symptomatically in the three years after screening), and therefore whether the intervention reduced the number of cancers missed at screening.

3. Aims and Objectives of the Trial

3.1 Primary aim

To determine whether reversing the order in which the second reader reads the mammograms affects cancer detection rate.

3.2 Secondary aims

To investigate the mechanism by which the intervention is hypothesised to increase cancer detection rate. This is made up of several parts:

1. What effect does the intervention have on recall rate?
2. How well does the arbitration process perform?
3. Does reversing the order in which the second reader reads the mammograms increase disagreement between readers
4. Does cancer detection rate and recall rate change with time since a break?

To determine whether reversing the order in which the second reader reads the mammograms affects missed cancer rate.

3.3 Outcome Measures

a) Efficacy:
i) Primary

Cancer detection rate in intervention and control group.

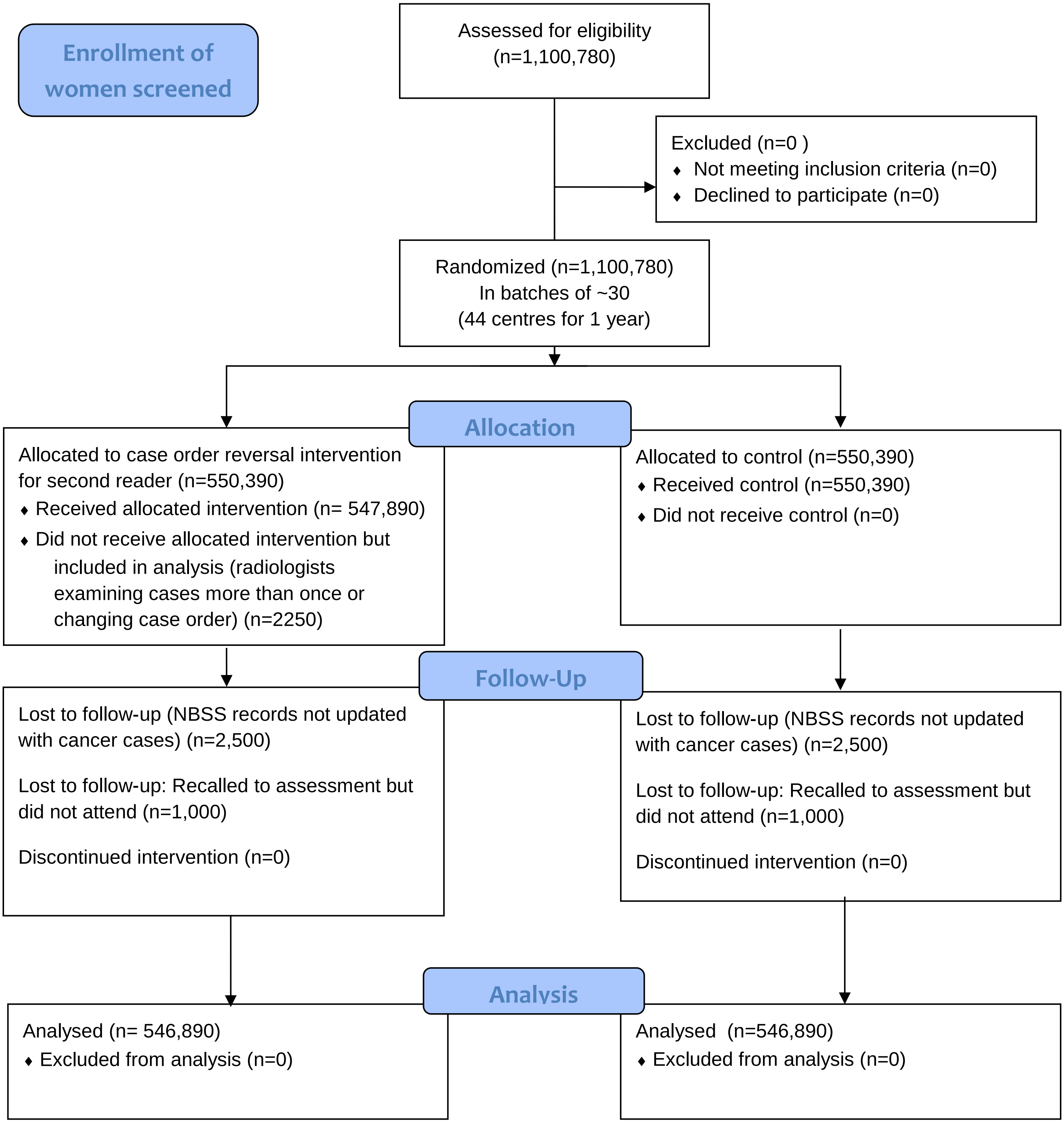
ii) Secondary

1. Recall rate in intervention and control groups
2. Referral rate to arbitration, recall rate and cancer detection rate from arbitration, and positive predictive value of arbitration by type of arbitration used. Sensitivity and specificity of arbitration using 3 years follow up data
3. Disagreement rate in intervention and control group
4. Change in disagreement rate, recall rate, cancer detection rate, and missed cancer rate with intended reading order (and cases since a cancer, recall, and mid-batch break)
5. Rate of cancers detected symptomatically in the three years subsequent to screening in intervention and control group. To be collected three years after the trial implementation is complete. This provides a measure of the difference in rates of cancer missed.

iii) Health economics
Cost per extra cancer detected
Cost per quality adjusted life year gained from the intervention

2.2 Trial flow diagram

CONSORT 2010 Flow Diagram for cancer detection analysis after 12 months



Assessed for eligibility (n=1,100,780)

Excluded (n=0)
- Not meeting inclusion criteria (n=0)
- Declined to participate (n=0)

Randomized (n=1,100,780)
In batches of ~30
(44 centres for 1 year)

Allocated to case order reversal intervention for second reader (n=550,390)
- Received allocated intervention (n= 547,890)
- Did not receive allocated intervention but included in analysis (radiologists examining cases more than once or changing case order) (n=2250)

Allocated to control (n=550,390)
- Received control (n=550,390)
- Did not receive control (n=0)

Lost to follow-up (NBSS records not updated with cancer cases) (n=2,500)

Lost to follow-up: Recalled to assessment but did not attend (n=1,000)

Discontinued intervention (n=0)

Lost to follow-up (NBSS records not updated with cancer cases) (n=2,500)

Lost to follow-up: Recalled to assessment but did not attend (n=1,000)

Discontinued intervention (n=0)

Analysed (n= 546,890)
- Excluded from analysis (n=0)

Analysed (n=546,890)
- Excluded from analysis (n=0)

3.4 Statistical Analysis Plan

i) Primary

To determine whether cancer detection rate is higher in the intervention group in comparison to the control group analysis will be conducted using a multi-level logistic regression model.

Levels to be included in the model are case, batch, and centre. To prevent over-fitting each level will only be included if they explain a sufficient portion of the variability.

It is reasonable to expect characteristics at the level of the women screened to be very similar between intervention and control due to the large number of batches of women randomised. Similarly centre level characteristics such as the arbitration system used will not differ between intervention and control as randomisation is at the batch level within each centre. Therefore the main analysis will simply contain one predictor, intervention or control arm. The intervention will be considered effective if the 5th percentile of the distribution of the coefficient for the intervention in the model is greater than zero.

A second analysis will be conducted including adjustment for characteristics of the woman screened (age and whether they have previously attended screening). Sub-group analysis will be conducted in younger women, those whose cases are read at the beginning or end of the batch, and younger women, as the intervention may be more effective in these groups.

ii) Secondary

Analysis of recall rate and rate of disagreements between intervention and control group will use the same methods as for the primary analysis. The positive predictive value of intervention and control group will also be calculated.

4. Power and Sample Size

The trial is powered to detect at the 5% significance level with 80% power in a two tailed test a difference in cancer detection rate between intervention and control group of one extra cancer detected per 2000 women screened. This would correspond to around 1000 extra cancers detected by screening in England each year. Power calculations used
cancer detection rate in the control group (7.8 per thousand women screened) and cancer detection rate in the intervention group (8.3 per thousand women screened). These estimates of cancer detection rate are derived from data from a previous study.

ICC was calculated from previous data, using a logistic binomial-Gaussian model (method B) with 1000 Monte Carlo simulations.[7] Hence, using the derived ICC of 0.002 and a cluster (batch) size of 40 women, this then gives the design effect as 1.09. This gave an overall sample size required of 1,093,780 women, or 44 breast screening centres for 1 year (On the basis that in England there are 82 centres each screening around 25,000 women per year).

There is no adjustment for drop out or crossover because once the intervention is applied to a screening centre each batch will automatically be randomly assigned to intervention or control groups by the NBSS computer system, and the intervention applied automatically by that same system. However, there is a possibility that a small minority of records (circa 5000) will not be updated in time for the data collection. These will be considered missing at random as they will be all the cases from certain dates. A further 2000 women may be recalled for further tests and not attend their appointment. Furthermore, some radiologists may not read cases in the intended order, for example coming back to re-read difficult cases at a later time. This also is expected to be very rare in a busy screening centre, (circa 2250 in intervention arm), these cases will be included in the analysis. Therefore 1,100,780 cases are required to be randomised, equivalent to 44 screening centres for 1 year.

The same method was applied to detect a difference in cancer detection rate between intervention and control group of one fewer cancer missed per 2500 women screened. This gave an overall sample size required of 1.5 million women, or 44 breast screening centres for 1 year and 4 months.

Therefore 44 breast screening centres will be recruited to the trial, which will last for 1 year and 4 months. After the first year of the trial (allowing 2 months for follow up appointments to determine which women have cancer) the data for one year of screening at 44 centres will be downloaded, and analysis of the primary outcome of cancer detection rate, and the secondary outcomes to be completed. Analysis of missed cancer rates will not be until three years after completion of the whole trial duration.

There are 82 Breast Screening centres in England, of which 48 are fully digital, and 25 are partly digital. It is anticipated that by the start of data collection 64 centres will be fully digital. It is anticipated that it will take three months to enrol 44 centres and obtain local R&D approvals. Administering the trial at each centre will take the NHS staff no extra time commitment as it is fully computerised and will be automatically downloaded with their latest software update, which is anticipated will help reach recruitment targets.

5. Eligibility

5.1 Inclusion criteria

Screening centres taking part must use double reading of screening mammograms, and must have at least one piece of digital mammography equipment used for screening

All women who receive mammography screening using digital equipment during the study period at the study centres.

5.2 Exclusion criteria

Centres which use single reading of mammograms

Centres which have no digital mammography equipment used for screening (the software intervention does not work on the old equipment)

Women who attend symptomatic breast clinics

5.3 Post-randomisation withdrawals and exclusions

Centres participating may withdraw from the trial treatment, and/or the whole trial at any time without prejudice. No follow up data will be collected, only the initial survey results describing the centre characteristics will be used to analyse whether there are patterns in withdrawals from the trial, unless the centre requests that their data is not used in this way.

5.4 Compliance/contamination

Compliance is expected to be very high, as after centre level consent is given all implementation is automated. Therefore deviation from the protocol would be very difficult. Similarly contamination is not anticipated to be an issue as there is no method available to move cases between the intervention and control groups. However case order can be changed by selecting the ignore option rather than inputting a screening result and reviewed at a later stage, or a result could be inputted and then revised later. These things would change the reading order and produce some level of contamination. This will be ameliorated by analysing based on intended order of reading. Data will be collected on actual reading order also to measure levels of contamination and their effect on outcomes. These effects are expected to be extremely small (of the order of once per 500 cases) because screening is a fast moving high volume activity and there is no time in practice to come back to cases at a later stage.

6. Consent

Informed consent will be obtained at the centre level, by the director of the service and the lead radiologist. These are usually the same person but if they are not the consent of both will be sought. Consent will be required at the centre level rather than the patient or radiologist as it is at this level that the intervention is applied. The intervention can be considered an alternative and at least equivalently good form of standard practice, as nothing about how the mammograms are reviewed and evaluated changes, just the order in which they are assessed.

Directors of breast screening centres will be contacted in the first instance by email introducing the study with a copy of the participant information sheet and informed consent form attached. They will then receive a follow up phone call one week later. Each centre will be offered a copy of the research protocol and IRAS ethics form, and a visit from one of the investigation team to introduce the study if they wish. The study software which will run the trial will be embedded in the NBSS software in every English breast screening as part of the routine updating of the NBSS system. This software will remain inactive until a signed consent form is received from that centre, at which point it can be activated simply and quickly. The study can commence in each centre by simply changing the software settings to activate the intervention.

In the unlikely event that information becomes available may be relevant to the participant’s willingness to continue in the trial, for example findings from other research studies, then the directors of every participating breast screening centre will be informed immediately by email and follow up phone call.

6.1 Recruitment and Randomisation

The rate of accrual will be monitored at the centre level, if it falls appreciably below the projected level, the reasons will be identified and remedial actions taken in order to protect the power of the trial and alleviate concerns about selective entry and other aspects of quality.

Randomisation will be automatically computer generated at the point at which the batch is ready to be read on the NBSS system. The random number generation will be using the root of the time at which randomisation is required. Only batches of cases to be read as part of the NHS breast screening programme will be randomised. Cases from symptomatic clinics will be excluded prior to randomisation.

6.2 Blinding

There is no expected placebo effect in this trial. No elements of standard practice are to be changed as part of the trial except reading order, to test whether the intervention would be effective in normal practice. The women screened will not be aware whether their mammograms are read as part of an intervention or control batch. The radiologists and radiography advanced practitioners when acting as second reader are not blinded to reading order, but they are blinded to trial arm because both intervention and control groups include reading forwards and backwards. The statistician will be blinded to treatment groups until analysis is complete.

7. Project Timetable and Milestones

Year 1

August 2012: Software developed, ethical approvals complete and trial registered
December 2012: Pilot complete and 44 sites recruited

Year 2

Implementation in all 44 sites throughout 2013

Year 3

May 2014: 1 year data collected in 44 centres.
September 2014: Write up complete
December 2014: Heath economics analysis complete

Year 6

Collection and analysis of interval cancer data. Follow on funding is required to proceed with this part of the project.

8. Protocol Amendments

Version 2.0: agreed on 24th July 2013 – number of centres increased from 36 to 44. (Coventry and Warwickshire REC on 12th July 2013 approved minor amendment to increase the number of centres from 36 to 44).

9. References

1. Office of National Statistics, Mortality Statistics: Deaths registered in England and Wales, 2010, [accessed 23rd May 2012] Available from: http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-230730
2. Cancer Reform Strategy. London: The Department of Health, 2007 [accessed 23rd may 2012]. Available from: http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/documents/digitalasset/dh_081007.pdf
3. Breast Screening Programme, England 2010-11. Leeds: The NHS Information Centre, 2012 [accessed 23rd May 2012]. Available from: http://www.ic.nhs.uk/statistics-and-data-collections/screening/breast-screening/breast-screening-programme--england-2010-2011
4. Beral V, Cush S, Ellis IO, Emery J, Faulkner K, Given-Wilson R, et al. Screening for breast cancer in England: Past and future. Sheffield: NHS Cancer Screening Programmes, 2006. Report No.:61
5. Taylor-Phillips S, Wallis MG, Gale AG. Should previous mammograms be digitised in the transition to digital mammography? Eur Radiol 2009;19(8):1890-6.
6. See JE, Howe SR, Warm JS, Dember WN. Metaanalysis of the Sensitivity Decrement in Vigilance. Psychol Bull 1995;117(2):230-49.
7. Goldstein H, Browne W, Rasbash J: Partitioning variation in multilevel models. Underst Stat 2002, 1:223–231. doi:10.1207/S15328031US0104_02.

Supplement 2: Statistical Analysis Plan

Changing Case Order to Optimise Patterns of Performance in Screening (CO-OPS)
Randomised Controlled Trial

STATISTICAL ANALYSIS PLAN

Contents

1. Dates of amendments and decisions

24/10/12 ISRCTN application submitted

26/3/2013 ISRCTN application accepted (ISRCTN46603370)

24/7/2013 Statistical Analysis plan agreed by Trial Steering Committee

10/1/2014 Protocol published (Taylor-Phillips et al. 2014)

16/1/2014 Planned reporting of additional details of clusters added (table 1 added) to describe external validity (Eldridge et al. 2008)

20/3/2014 Agreement at Trial Steering Committee to add analysis 3. Agreement that this is exploratory follow-up analysis which will not form part of the analysis in the main paper

2/4/2014 Trainees at one participating centre are logging in as qualified film readers. This causes a subset of cases intended for intervention arm 1 to move to intervention arm 2. Decision by Trial Steering Committee not to change analysis as they remain in the intervention arm.

15/9/2014 Page 4 single reader, >2 readers, barcode reading, and administrator entering results misclassified 'protocol violation' corrected to 'non-compliance'.

23/10/2014 First data collected from breast screening centres

2. Power and Sample Size

Currently, 14,700 cancers are detected by screening each year and the cancer detection rate is 7.8 per thousand women screened. (NHS Information Centre, 2012) To detect one extra cancer per 2000 women screened, this would increase the cancer detection rate to around 8.3 cancers per 1000 women screened. To detect this change of 0.5 cancers per 1000 women, for a two-tailed test at 80% power and 5% significance, 501,361 women are required in each arm, see figure 1.

However, as randomisation occurs at the batch level, collected data is clustered and must also be taken into consideration. The sample size of a clustered study must be increased by the Design Effect (DE), which is calculated as $DE = 1 + (m - 1)\rho$ for a given ICC (ρ) and cluster size (m).

ICC was calculated from previous data, using a logistic binomial-Gaussian model (method B) with 1000 Monte Carlo simulations. (Goldstein, 2002) Hence, using the derived ICC of 0.002 and a cluster (batch) size of 40 women, this then gives the DE as 1.09.

Therefore, the overall sample size required is for 1,093,780 women, or 44 breast screening centres for 1 year (On the basis that in England there are 82 centres each screening around 25,000 women per year).

There is no adjustment for drop out or crossover because once the intervention is applied to a screening centre each batch will automatically be randomly assigned to intervention or control groups by the NBSS computer system, and the intervention applied automatically by that same system. A woman could be lost to follow up if she is recalled from screening for further tests and does not attend her follow up appointment, however this is uncommon and so we have assumed low dropout rates for those who are recalled for further tests.

3. Data extraction and models

One year of data will be extracted from each of the 46 centres taking part using an NBSS report developed by Sue Hudson at Acamdex, based on the trial tables in NBSS developed by Temenos Ltd. Data extraction will be 8 weeks after the one year is complete, to allow time for women to be recalled for further tests and the results entered into the database. Levels of missing data will be assessed, and a follow up data extraction may be performed. This would be achieved by re-extracting the entire dataset.

One line of data will be extracted for each woman who attended screening during the data collection period.

The following data items will be collected for each woman. Italicised text in square brackets indicates NBSS variables that will be extracted.

Possible levels in model

The following variables will be used to construct a multi-level model (see section 6 for more details).

- Centre ID
- Batch ID
- Woman ID

Model predictors from NBSS

The following variables will be used to construct a multi-level model (see section 6 for more details).

- Trial arm: Intervention or Control *[TrialArm Intervention =FR and RF, Control = FF and RR]* Trial assistant to blind the dataset before analysis by STP. TrialArm to be replaced by condition 1 and condition 2. Case order columns to be removed, leaving only whether each case was read in the intended order.
- Woman's age at screening
- Prevalent or incident screen

Subgroups for subgroup analysis

The following variables will be used to construct subgroups for analysis (see section 8 for more details).

- Woman's age in groups: 52 and under, 53-59, 60 and over
- Case position in batch *[number of cases from either the beginning or end of the batch, whichever is smaller]*, with extra column dichotomising into 'first or last five' and 'other'.
- Whether the batch was read first in a workday by both readers *[calculated by whether another batch was read by the same reader that day for both readers]*

Model outcomes

The following variables will be used as model outcomes (see section 6 and 7 for more details).

- Primary: Cancer detected at screen yes / no
- Secondary: Recall Yes/No
- Secondary: Disagreement between readers Yes/No
- Secondary: Health economics: Separate extract

Details of missing data

- The following variables will be used establish the causation of missing data. Recalled women who did not attend their recall appointment
- Whether the results of her recall appointment/surgery are available
- Missing data counts for all variables considered for inclusion in the model

Quality assurance/contamination/protocol violations

Analyses will be conducted as intention to treat unless specified. However, quality assurance measures will also be collected such as contamination of the intervention and protocol violations.

- Contamination: Proportion of cases read in a different order to that intended in the protocol, caused for example by readers moving cases to the end of a batch.
- Contamination: Reader did not complete whole batch in one session. When one reader stops reading batch half way through and come back to it later.
- Non-compliance: Only one reader and not a technical recall
- Non-compliance: More than two readers *(means that a trainee has logged in as a real reader, or there were 3 reads)*
- Non-compliance: Digital mammography batches or cases read using barcode so not part of the trial
- Non-compliance: Result entered by administrator rather than reader so order is not known
- Non-compliance: Trainee or administrator logs in as a reader and enters results in R1 or R2

Exclusions

The following cases will be excluded from the analysis

- Technical recall
- Same woman screened twice in 1 year, use only the first screening appointment

Other data collection variables

Other to add to database for future exploratory analysis:

- Fully blinded yes/No (from survey of breast screening centres)
- Arbitration method: 1, 2, 3ormore readers, arbitrate all or disagreements, same people read and arbitrate (from survey of breast screening centres)
- Method of displaying priors (digitise / hang/ available next to workstation) (from survey of breast screening centres)
- Centre size (total number of women in trial in 1 year)
- Number of cases in batch

A separate data extraction will be performed after 3 years to analyse the secondary outcome of interval cancer rate.

4. Data Cleaning

The readers in the NBSS extract will be checked to ensure they are independent readers and not the same person with 2 logins. The number of readers in the NBSS extract will be checked against the number of readers in the survey, and any discrepancies corrected in the survey (including reader type).

A small minority of women may have been screened twice during the data collection period, due to administrative errors, moving GP and rescreen in error, or other rare anomalous reasons. These cases will be identified.

Technical recalls will be counted and removed, missing data, QA and non-compliance will be recoded as per above.

5. Descriptive statistics

Table 1. Characteristics of centres taking part in comparison to those not taking part

	Centres taking part (n=47)	Centres not taking part (n=33)
Mean number of women screened in one year (sd)		
Mean percentage uptake (sd)		
Mean cancer detection rate (sd)		
Region: North East North West Yorkshire and the Humber East Midlands West Midlands East of England London South East Coast South Central South West		

X cases were included in the trial. These were examined by X qualified readers, of which X were radiologists, Y were radiography advanced practitioners and Z were breast clinicians. The mean batch length was X cases (standard deviation y cases)

In X centres arbitration is only when reader 1 and 2 disagree, but in Y centres all recalled cases are arbitrated, even if both reader agree the case should be recalled. A whole range of arbitration methods are used, X centres use one extra reader (third reader arbitration), Y centres use two reader consensus, and Z centres use 3 or more reader consensus.

In X centres the second reader cannot see the first readers decision, in Y centres they cannot see on the computer software but can by looking at the paperwork, and in Z centres the first readers decision is visible onscreen.

All centres taking part in the trial used digital mammography for screening. In X centres the mammograms from the previous screening round were displayed digitally, in Y centres the films were pre-hung on a multiviewer adjacent to the workstation, and in Z centres the film mammograms were available in a screening bag adjacent to the workstation for the readers to hang themselves on a light box if they wished.

Figure 1 Consort diagram

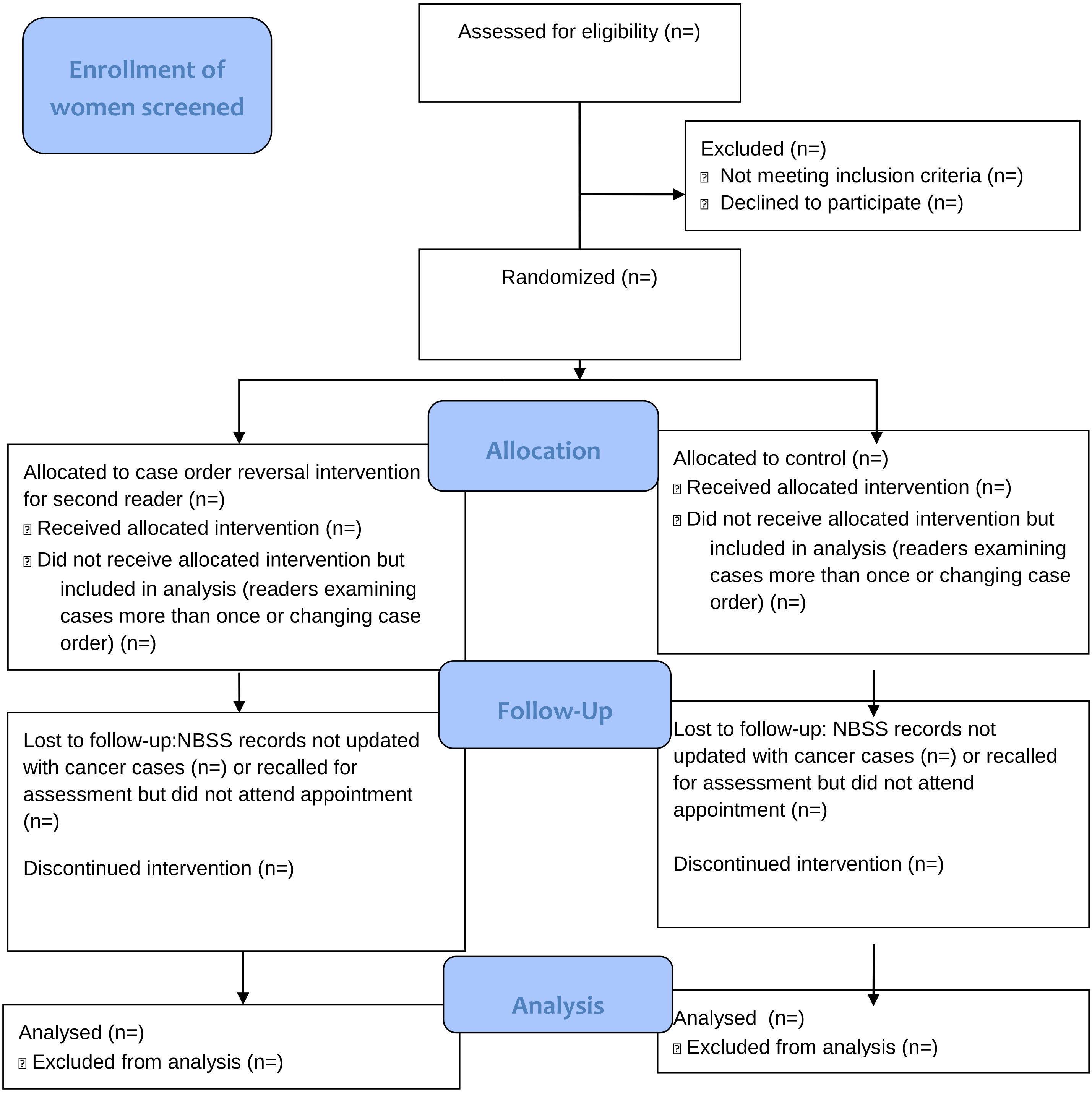

Enrollment of women screened
Assessed for eligibility (n=)
Excluded (n=)
□ Not meeting inclusion criteria (n=)
□ Declined to participate (n=)
Randomized (n=)
Allocation
Allocated to case order reversal intervention for second reader (n=)
□ Received allocated intervention (n=)
□ Did not receive allocated intervention but included in analysis (readers examining cases more than once or changing case order) (n=)
Allocated to control (n=)
□ Received allocated intervention (n=)
□ Did not receive allocated intervention but included in analysis (readers examining cases more than once or changing case order) (n=)
Follow-Up
Lost to follow-up:NBSS records not updated with cancer cases (n=) or recalled for assessment but did not attend appointment (n=)
Discontinued intervention (n=)
Lost to follow-up: NBSS records not updated with cancer cases (n=) or recalled for assessment but did not attend appointment (n=)
Discontinued intervention (n=)
Analysis
Analysed (n=)
□ Excluded from analysis (n=)
Analysed (n=)
□ Excluded from analysis (n=)

Table 2 Descriptive statistics for intervention and control groups (NB these will be group 1 and group 2 until the statistician is unblinded)

	Intervention	Control
Mean age of women screened (sd)		
Mean batch length (sd)		
Cancer detection rate (95%CI)		
Recall rate (95%CI)		
Disagreement rate (95%CI)		

Table 3 Descriptive statistics for reading forwards or backwards

	Forwards	Backwards
Mean time taken per case (sd)		
Reader recall rate (sd)		

6. Statistical Analysis for primary outcome

To determine whether cancer detection rate is higher in the intervention group in comparison to the control group a two-tailed analysis will be conducted using a multi-level logistic regression model in Stata calling MLwiN software. MlwiN will be used because it can compute multi-level models for very large datasets without excessive computer hardware requirements.

Two models will be constructed as detailed below. For all models residuals will be examined for outliers. Analysis will be conducted as intention to treat, with all cases randomised included in the analysis. Missing data through loss to follow up will occur in both groups. This will include women who have been recalled from screening, but either did not attend their follow up appointment or there are no records in the database concerning the results of that appointment. Multiple imputation and sensitivity analysis to examine the effects of any missing data on the model will be considered if missing data is above 5% of the data set. Otherwise complete case analysis will be used.

Analysis 1:

The first model will include only treatment as a predictor of cancer detection. Levels considered for inclusion in the model will be: case; batch; and centre. To prevent over-fitting, each level will only be included in the final model if it explains a sufficient portion of the variability and improves model fit. Screening centre will be added as a level to the model to account for clustering. It will be retained in the model if it improves model fit (Wald test at the 5% level, see limitations section). Then batch will be considered for addition as a level to this model if it improves model fit according to the same criteria. The treatment will be added to this model.

A sub-analysis of those cases which are intended to be read in the first or last 5 of the batch, younger women (≤52 , 53-59, ≥60) and the first batches to be read in a workday (by both readers) will be conducted.

Analysis 2:

The second model will adjust for other predictors of cancer detection in order to correct for any imbalances in these between intervention and control group.

First a model will be constructed with known predictors of cancer detection rate. The woman’s age and whether she has previously attended screening are known to affect cancer detection rate and so will be included in the model as fixed effects. Then screening centre and batch will be tested for inclusion as levels in the model, using the same process as analysis 1. The treatment will be added to this baseline model.

Analysis 3:

This is exploratory follow-up analysis but will not form part of the analysis in the main paper

The third model will be used to determine whether there are particular circumstances in which the intervention is effective.

Predictors to be considered for inclusion will be added separately as fixed effects to the baseline model from analysis 2. Those predictors which improve the model (Wald statistic sig at 10% level) will be added to one model, with those with highest Wald statistic added first. Those for which the Wald statistic remains significant at the 10% level will be retained in the baseline model. Predictors to be considered for inclusion are number of readers involved in arbitration (categorical: 1/2/3 or more), arbitration policy (binary: all recalled cases arbitrated/disagreements only arbitrated), arbitration independence (same/ different readers read and arbitrate the case) whether reader 2 can see reader 1 opinion (categorical: yes on computer/yes but only on paperwork/no), how the mammograms from the previous screening round are displayed (categorical: digital/film pre-hung on a multi-viewer/film available in bag by the workstation), whether the batch was read immediately after another batch(categorical no/yes one reader/yes both readers), the number of cases in the batch, and where in the batch the case was intended to be read (number of cases from beginning/end).

The treatment will be added to this model. The interaction between treatment (intervention or control) and each of these predictors will then be considered for inclusion using the same process and criteria.

7. Statistical analysis for secondary outcomes

To determine whether the number of disagreements between readers, interval cancer rate, and recall rate is different between the intervention and control groups (secondary aim i, ii, and iii); the same methods will be used as described for the primary analysis above.

The Positive predictive value (PPV) of cancer detection in each study arm will be calculated as the proportion of women recalled who are found to have cancer (secondary aim iii). The difference between PPV in the control and intervention arms will be investigated using the same methods as for the primary outcome, but including only recalled cases.

The effects of the trial introducing the reverse reading order (secondary aim v) will also be analysed. Here, the recall and cancer detection rates for the two reading groups which make up the control arm will be compared. Models will be constructed as for the primary analysis, but using reading order as the predictor of cancer detection rate and recall rate, not trial group membership.

To generate secondary outcome 4 (estimates of cost-effectiveness), the primary outcome from the trial will be used as an input into a health economic model of breast cancer screening. This model will be developed by expanding an earlier breast cancer model by Campbell et al. (2011) and will predict lifetime costs and effects for both intervention and control arms.

8. Other analyses with same data set

- Performance with number of cases since the beginning of the batch. *[R1intendedOrder]*
- Effects on performance of reader completing one session after another and with time of day *[Calculate batch in day by assigning a clinic start time for each reader using date and time stamp for first case read for each ReportingSetID]*.

9. References

Campbell HE, Epstein D, Bloomfield D, Griffin S, Manca A, Yarnold J, Bliss J, Johnson L, Earl H, Poole C, Hiller L, Dunn J, Hopwood P, Barrett-Lee P, Ellis P, Cameron D, Harris AL, Gray AM, Sculpher MJ: The cost-effectiveness of adjuvant chemotherapy for early breast cancer: a comparison of no chemotherapy and first, second, and third generation regimens for patients with differing prognoses. Eur J Cancer 2011, 47:2517–2530.

Eldridge S, Ashby D, Bennett C, Wakelin M, Feder G. Internal and external validity of cluster randomised trials: systematic review of recent trials. *BMJ : British Medical Journal*. 2008;336(7649):876-880. doi:10.1136/bmj.39517.495764.25.

Golstein H, Browne W, Rasbash J: Partitioning variation in multilevel models. Underst Stat 2002, 1:223–231. doi:10.1207/S15328031US0104_02.

Taylor-Phillips S, Wallis MG, Parsons H, Dunn J, Stallard N, Campbell H, Sellars S, Szczepura A, Gates S, Clarke A. Changing case Order to Optimise patterns of Performance in mammography Screening (CO-OPS): study protocol for a randomized controlled trial. Trials. 2014 Jan 10;15:17. doi: 10.1186/1745-6215-15-17.

The NHS Information Centre, Screening and Immunisations: Breast Screening Programme, England—2010–11 [NS]. Leeds, UK: The Health and Social Care Information Centre; 2012. Available at https://catalogue.ic.nhs.uk/publications/screening/breast/bres-screprog- eng-2010-11/bres-scre-prog-eng-2010-11-rep.pdf (accessed 16 December 2013).

Supplement 3

A. Characteristics of Participating Centres

There are 80 breast screening centres in England, each centre represents a single or group of hospitals in a region. Of these, 46 centres agreed to take part, we were unable to contact the director of breast screening at 22 centres, six centres were interested but had equipment incompatible with the trial software, two did not want to take part as they already used the intervention for all cases, and four were simply not interested. The characteristics of centres taking part in comparison to other centres in England is detailed in table e1. One centre withdrew early from the trial due to practical difficulties experienced with the change to case order, data from their first four months is included in the analysis.

Table e1. Characteristics of centres taking part in comparison to those not taking part. Data for the year 2011-12

	Centres taking part (n=46)	Centres not taking part (n=34)
Mean number of women screened in one year (sd)	21,921 (10,355)	18,229 (10,349)
Mean percentage uptake[‡] (sd)	74.0 (3.3)	73.9 (5.6)
Mean percentage recall (sd)	3.62 (1.24)	3.47 (1.23)
Mean small (<15mm) cancer detection rate per thousand women screened who have previously attended screening (sd)	3.2 (0.7)	3.4 (0.7)
Standardised detection rate incident round cancers (sd)	1.47 (0.21)	1.49 (0.22)
Region:		
North East	1	3
North West	10	1
Yorkshire and the Humber	6	2
East Midlands	2	6
West Midlands	6	2
East of England	5	6
London	1	5
South East Coast	5	1
South Central	3	6
South West	7	2

[‡] The proportion of invited women who attend screening, expressed as a percentage

B. Detailed Modelling Results

Multilevel analysis for cancer detection models

Measures of association

Table e2 shows results of multilevel models for case, batch and centre level factors associated with breast cancer detection rates in the UK. With all factors controlled for in the multilevel analysis, age of participants and having first ever screen were statistically significantly associated with the odds of breast cancer detection. Every one year increase in the age of participants increased the odds of cancer detection by 5% (OR 1.05; 95% CI 1.04 - 1.06, p<0.001). For participants who have never been screened before the odds of cancer detection were 73% greater than for those participants who have been screened before. When all the factors were controlled for in the final model, the odds of a participant being detected as having breast cancer if in the treatment arm was very similar to that in the control arm (OR 1.01; 95% CI 0.97 – 1.06, p=0.660).

Measures of variation

Table e2 shows random-effect (measures of variation) results from the multilevel analysis. In Model 1 (the empty model), there was a significant variation in the log odds of cancer detection across the batches (τ = 0.811, 95% CI 0.757 – 0.866) and across the centres (τ = 0.058, 95% CI 0.012 – 0.104). According to the intra-centre and intra-batch correlation coefficient implied by the estimated intercept component variance, 1.4% and 20.9% of the variance of cancer detection could be linked to the centre- and batch-level factors, respectively. Variations across batches and centres remained statistically significant, even after controlling for treatment and background factors in the final model 4, thereby giving credence to the use of multilevel modelling to account for batch and centre variations.

The median odds ratio (MOR) results also confirmed the evidence of batch and centre contextual phenomena modifying the likelihood of cancer detection. In Models 1 and 2 the batch level heterogeneity is high (MOR of 2.35), but the centre level heterogeneity is low (MOR of 1.26). When the model is adjusted for age and whether the woman has been screened before (Models 3 and 4) the batch level MOR (2.08) remains high and the centre level MOR (1.20) remains low.

Table e2: Factors associated with cancer detection rate identified by multilevel multivariable logistic regression models.

Variable	Model 1[a] **OR (CI)**	Model 2[b] **OR (CI)**	Model 3[c] **OR (CI)**	Model 4[d] **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)	1.052(1.048-1.055)
No previous attendance			1.73(1.62-1.86)	1.73(1.62-1.86)
RANDOM-EFFECTS (measures of variation)				
Centre level				
Variance (SE)	0.058(0.012-0.104)	0.058(0.012-0.104)	0.038(0.011-0.065)	0.038(0.011-0.064)
Intra-centre correlation (%)	1.40	1.39	0.96	0.96
MOR	1.26	1.26	1.20	1.20
Wald statistics (p-value)	0.014	0.014	0.006	0.006
Batch level				
Variance (SE)	0.811 (0.757-0.866)	0.809(0.754-0.863)	0.598(0.546-0.650)	0.595(0.543-0.647)
Intra-batch correlation (%)	20.90	20.85	16.19	16.13
MOR	2.35	2.35	2.08	2.08
Wald statistics (p-value)	<0.001	<0.001	<0.001	<0.001

[a]Model 1 is the empty model, a baseline model without any predictor variable
[b]Model 2 is adjusted for treatment variable
[c]Model 3 is adjusted for background factors (age and previous attendance)
[d]Model 4 is adjusted for treatment and background variables (age and previous attendance)
Abbreviations: SE; standard error, CI; confidence interval, MOR; median odds ratio.

Multilevel analysis for recall rate models

Measures of association

Table e3 shows results of multilevel models for case, batch and centre level factors associated with recall rates in the UK. With all factors controlled for in the multilevel analysis, age of participants and having first ever screen were statistically significantly associated with the odds of recall. Every one year increase in the age of participants increased the odds of recall by 0.8% (OR 1.008; 95% CI 1.007 - 1.010, p<0.001). For participants who have never been screened before the odds of being recalled for re-evaluation were 189% greater than those for participants who have been screened before. When all the factors were controlled for in the final model, the odds of a participant being recalled for re-evaluation if in the treatment arm (OR 0.997; 95% CI 0.978 – 1.016, p=0.726) decreased by 0.3% compared to those in the control arm, though this association was not statistically significant.

Measures of variation

Table e3 shows random-effect (measures of variation) results from the multilevel analysis. In Model 1 (the empty model), there was a significant variation in the log odds of cancer detection recall across the batches (τ = 0.104, 95% CI 0.092 – 0.116) and across the centres (τ = 0.052, 95% CI 0.030 – 0.074). According to the intra-centre and intra-batch correlation coefficient implied by the estimated intercept component variance, 1.5% and 4.5% of the variance of cancer detection recall could be linked to the centre- and batch-level factors, respectively. Variations across batches and centres remained statistically significant, even after controlling for treatment and background factors in the final model 4.

The MOR (1.36) in Model 1 across the batches suggests that batch heterogeneity is moderate. Controlling for treatment-level factor (Model 2) did not change the unexplained heterogeneity between batches. However, at the centre level the clustering effect is low (MOR of 1.24). The unexplained centre heterogeneity remained unchanged when all the factors were controlled for (Model 4). Thus, there were little variations between centres in the likelihood of recall.

Table e3: Factors associated with recall rate identified by multilevel multivariable logistic regression models.

Variable	Model 1[a] **OR (CI)**	Model 2[b] **OR (CI)**	Model 3[c] **OR (CI)**	Model 4[d] **OR (CI)**
FIXED-EFFECTS (measures of association)				
Treatment variable				
Treatment (vs. control)		0.993(0.974-1.013)		0.997(0.978-1.016)
Background factors				
Age (per year of age)			1.008(1.007-1.010)	1.008(1.007-1.010)
No previous attendance			2.89(2.82-2.97)	2.89(2.82-2.97)
RANDOM-EFFECTS (measures of variation)				
Centre level				
Variance (SE)	0.052(0.030-0.074)	0.052(0.030-0.074)	0.053(0.030-0.076)	0.053(0.030-0.076)
Intra-centre correlation (%)	1.51	1.51	1.57	1.57
MOR	1.24	1.24	1.24	1.24
Wald statistics (p-value)	<0.001	<0.001	<0.001	<0.001
Batch level				
Variance (SE)	0.104(0.092-0.116)	0.104(0.092-0.116)	0.033(0.023-0.044)	0.033(0.023-0.044)
Intra-batch correlation (%)	4.54	4.54	2.55	2.55
MOR	1.36	1.36	1.19	1.19
Wald statistics (p-value)	<0.001	<0.001	<0.001	<0.001

[a]Model 1 is the empty model, a baseline model without any predictor variable
[b]Model 2 is adjusted for treatment variable
[c]Model 3 is adjusted for background factors (age and previous attendance)
[d]Model 4 is adjusted for treatment and background variables (age and previous attendance)
Abbreviations: SE; standard error, CI; confidence interval. MOR; median odds ratio

Multilevel analysis for disagreement rate models

Measures of association

Table e4 shows results of multilevel models for case, batch and centre level factors associated with disagreement rates. With all factors controlled for in the multilevel analysis, age of participants and having first ever screen were statistically significantly associated with the odds of disagreement rates amongst the readers. Every one year increase in the age of participants reduced the odds of disagreement by 0.6% (OR 0.994; 95% CI 0.992 - 0.996, p<0.001). The odds of disagreement in participants who have never been screened before were 117% higher than those for participants who have been screened before. When all the factors were controlled for in the final model, the odds of disagreement between readers in making cancer call for a participant in the treatment arm (OR 0.997; 95% CI 0.974 – 1.020, p=0.780 decreased by 0.3% compared to those in the control arm, though this association was not statistically significant.

Measures of variation

Table e4 shows random-effect (measures of variation) results from the multilevel analysis. In Model 1 (the empty model), there was a significant variation in the log odds of disagreement across the batches (τ = 0.270, 95% CI 0.252 – 0.287) and across the centres (τ = 0.106, 95% CI 0.061 – 0.151). According to the intra-centre and intra-batch correlation coefficient implied by the estimated intercept component variance, 1.4% and 10.3% of the variance of cancer detection disagreement rate could be linked to the centre- and batch-level factors, respectively. Variations across batches and centres remained statistically significant, even after controlling for treatment and background factors in the final model 4.

At the centre level the MOR was 1.36 for all four models, suggesting that the clustering effect is moderate. At the batch level the MOR was 1.64 for Models 1 and 2, suggesting that the batch heterogeneity is moderate; reducing to 1.56 when all of the factors were introduced (Model 4), which is also indicative of moderate heterogeneity.

Table e4: Factors associated with disagreement rate identified by multilevel multivariable logistic regression models.

Variable	Model 1[a] **OR (CI)**	Model 2[b] **OR (CI)**	Model 3[c] **OR (CI)**	Model 4[d] **OR (CI)**
FIXED-EFFECTS (measures of association)				
Treatment variable				
Treatment (vs. control)		0.994(0.971-1.019)		0.997(0.974-1.020)
Background factors				
Age (per year of age)			0.994(0.992-0.996)	0.994(0.992-0.996)
No previous attendance			2.17(2.11-2.24)	2.17(2.11-2.24)
RANDOM-EFFECTS (measures of variation)				
Centre level				
Variance (SE)	0.106(0.061-0.151)	0.106(0.061-0.151)	0.104(0.060-0.148)	0.104(0.060-0.148)
Intra-centre correlation (%)	2.89	2.89	2.88	2.88
MOR	1.36	1.36	1.36	1.36
Wald statistics (p-value)	<0.001	<0.001	<0.001	<0.001
Batch level				
Variance (SE)	0.270(0.252-0.287)	0.270(0.252-0.287)	0.216(0.200-0.233)	0.216(0.200-0.233)
Intra-batch correlation (%)	10.25	10.25	8.87	8.87
MOR	1.64	1.64	1.56	1.56
Wald statistics (p-value)	<0.001	<0.001	<0.001	<0.001

[a]Model 1 is the empty model, a baseline model without any predictor variable
[b]Model 2 is adjusted for treatment variable
[c]Model 3 is adjusted for background factors (age and previous attendance)
[d]Model 4 is adjusted for treatment and background variables (age and previous attendance)
Abbreviations: SE; standard error, CI; confidence interval, MOR; median odds ratio.

Models of cancer detection rate by reading order

Measures of association

Table e5 shows results of multilevel models for case, batch and centre level factors associated with breast cancer detection rates in the UK. With all factors controlled for in the multilevel analysis, age of participants and having first ever screen were statistically significantly associated with the odds of breast cancer detection. Every one year increase in the age of participants increased the odds of cancer detection by 5% (OR 1.053; 95% CI 1.049 - 1.506, p<0.001). For participants who have never been screened before the odds of having cancer detected were 79% greater than those for participants who have been screened before. When all the factors were controlled for in the final model the odds of a participant being detected as having breast cancer decreases by 0.013% (OR 0.99987; 95% CI 0.998 – 1.001, p=0.869) for each case further along in the batch.

Measures of variation

Table e5 shows random-effect (measures of variation) results from the multilevel analysis. In Model 1 (the empty model), there was a significant variation in the log odds of cancer detection across the batches (τ = 1.115, 95% CI 1.053 – 1.178) and across the centres (τ = 0.090, 95% CI 0.020 – 0.160). According to the intra-centre and intra-batch correlation coefficient implied by the estimated intercept component variance, 2.00% and 26.8% of the variance of cancer detection could be linked to the centre- and batch-level factors, respectively. Variations across batches and centres remained statistically significant, even after controlling for reading order and background factors in the final model 4.

In Models 1 and 2 the batch level heterogeneity is high (MOR of 2.73 and 2.72 respectively), but the centre level heterogeneity is low (MOR of 1.33 for both models). When the model is adjusted for age and whether the woman has been screened before (Models 3 and 4) the batch level MOR (2.24) remains high and the centre level MOR (1.23) remains low.

Table e5: Models of reader 1 cancer detection rate by reading order, using multilevel multivariable logistic regression models.

Variable	Model 1[a] **OR (CI)**	Model 2[b] **OR (CI)**	Model 3[c] **OR (CI)**	Model 4[d] **OR (CI)**
FIXED-EFFECTS (measures of association)				
Treatment variable				
Reader 1 intended order (per case)		0.99966 (0.998-1.001)		0.99987 (0.998-1.001)
Background factors				
Age (per year of age)			1.053(1.049-1.056)	1.053(1.049-1.056)
No previous attendance			1.79(1.67-1.93)	1.79(1.67-1.93)
RANDOM-EFFECTS (measures of variation)				
Centre level				
Variance (SE)	0.090(0.020-0.160)	0.090(0.020-0.160)	0.049(0.016-0.081)	0.049(0.017-0.081)
Intra-centre correlation (%)	2.00	2.00	1.20	1.21
MOR	1.33	1.33	1.23	1.23
Wald statistics (p-value)	<0.001	<0.001	<0.001	<0.001
Batch level				
Variance (SE)	1.115 (1.053-1.178)	1.112(1.050-1.175)	0.719(0.661-0.777)	0.719(0.661-0.777)
Intra-batch correlation (%)	26.81	26.77	18.93	18.93
MOR	2.73	2.72	2.24	2.24
Wald statistics (p-value)	<0.001	<0.001	<0.001	<0.001

[a]Model 1 is the empty model, a baseline model without any predictor variable
[b]Model 2 is adjusted for intended treatment order
[c]Model 3 is adjusted for background factors (age and previous attendance)
[d]Model 4 is adjusted for treatment and background variables (age and previous attendance)
Abbreviations: SE; standard error, CI; confidence interval, MOR; median odds ratio.

Models of recall rate by reading order

Measures of association

Table e6 shows results of multilevel models for case, batch and centre level factors associated with recall rates in the UK. With all factors controlled for in the multilevel analysis, age of participants and having first ever screen were statistically significantly associated with the odds of recall. Every one year increase in the age of participants increased the odds of recall by 0.4% (OR 1.004; 95% CI 1.003 - 1.006, p<0.001). For participants who have never been screened before the odds of being recalled for re-evaluation were 170% higher than those for participants who have been screened before. When all the factors were controlled for in the final model the odds of a participant being recalled decreases by 0.29% (OR 0.9971; 95% CI 0.9966 – 0.9977, p<0.001) for each case further along in the batch.

Measures of variation

Table e6 shows random-effect (measures of variation) results from the multilevel analysis. In Model 1 (the empty model), there was a significant variation in the log odds of cancer detection recall across the batches (τ = 0.137, 95% CI 0.126 – 0.147) and across the centres (τ = 0.077, 95% CI 0.044 – 0.110). According to the intra-centre and intra-batch correlation coefficient implied by the estimated intercept component variance, 2.20% and 6.11% of the variance of cancer detection recall could be linked to the centre- and batch-level factors, respectively. Variations across batches and centres remained statistically significant, even after controlling for reading order and background factors in the final model 4.

At the centre level the MOR (1.30) was the same for all four models, suggesting that batch heterogeneity is low. At the batch level, in Model 1 and 2 the MOR (1.42 and 1.41 respectively) indicated that heterogeneity was moderate, but was low in Model 3 and 4 (MOR of 1.31 for both models).

Table e6: Models of reader 1 recall rate by reading order, using multilevel multivariable logistic regression models.

Variable	Model 1[a] **OR (CI)**	Model 2[b] **OR (CI)**	Model 3[c] **OR (CI)**	Model 4[d] **OR (CI)**
FIXED-EFFECTS (measures of association)				
Treatment variable				
Reader 1 intended order (per case)		0.9951(0.9946-0.9957)		0.9971(0.9966-0.9977)
Background factors				
Age (per year of age)			1.005(1.003-1.006)	1.004(1.003-1.006)
No previous attendance			2.72(2.66-2.79)	2.70(2.64-2.77)
RANDOM-EFFECTS (measures of variation)				
Centre level				
Variance (SE)	0.077(0.044-0.110)	0.076(0.044-0.108)	0.077(0.044-0.110)	0.076(0.044-0.109)
Intra-centre correlation (%)	2.20	2.17	2.24	2.21
MOR	1.30	1.30	1.30	1.30
Wald statistics (p-value)	<0.001	<0.001	<0.001	<0.001
Batch level				
Variance (SE)	0.137(0.126-0.147)	0.132(0.121-0.142)	0.080(0.070-0.089)	0.080(0.070-0.089)
Intra-batch correlation (%)	6.11	5.94	4.55	4.53
MOR	1.42	1.41	1.31	1.31
Wald statistics (p-value)	<0.001	<0.001	<0.001	<0.001

[a] Model 1 is the empty model, a baseline model without any predictor variable

[b] Model 2 is adjusted for intended treatment order

[c] Model 3 is adjusted for background factors (age and previous attendance)

[d] Model 4 is adjusted for treatment and background variables (age and previous attendance)

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