Main Manuscript for: Fatigue and Vigilance in Medical Experts Detecting Breast Cancer

Journal title: Proceedings of the National Academy of Sciences (PNAS)

Authors: Sian Taylor Phillips[1]*, David Jenkinson[1], Chris Stinton[1], Melina A Kunar[2], Derrick G Watson[2], Karoline Freeman[1], Alice Mansbridge[1], Matthew G Wallis[3], Olive Kearins[4], Sue Hudson[5], Aileen Clarke[1]

Affiliations:

1. Division of Health Sciences, Warwick Medical School, University of Warwick; Coventry, CV4 7AL, UK

2. Department of Psychology, University of Warwick, Coventry, CV4 7AL, UK

3. Cambridge Breast Unit and NIHR Cambridge Biomedical Research Centre, Cambridge University Hospitals NHS Trust; Hills Road. Cambridge, CB2 0QQ, UK

4. Screening Quality Assurance Service, NHS England; 23 Stephenson St., Birmingham, B2 4HQ, UK

5. Peel & Schriek Consulting Limited; London, UK

*Corresponding author. Email: s.taylor-phillips@warwick.ac.uk

Author contributions: STP, AC and MW conceived the study and contributed to the study design. STP acquired study funding. STP and SH acquired the data. DJ performed the analyses. STP, DJ, MK, DGW, CS, AM, and KF contributed to the interpretation of results. All authors wrote and edited the manuscript.

Competing interest declaration: 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 and is employed by NHS England.

Classification: Social Sciences > Psychological and Cognitive Sciences

Keywords: Breast Cancer, Cancer screening, Human Behavior, Vigilance Decrement

This file includes:
Main Text
Figures

Abstract

An abundance of laboratory-based experiments has described a vigilance decrement of reducing accuracy to detect targets with time on task, but there are few real-world studies, none of which have previously controlled the environment to control for bias. We describe accuracy in clinical practice for 360 experts who examined >1 million women's mammograms for signs of cancer, whilst controlling for potential biases. The vigilance decrement pattern was not observed. Instead, test accuracy improved over time, through a reduction in false alarms and an increase in speed, with no significant change in sensitivity. The multiple decision model explains why experts miss targets in low prevalence settings through a change in decision threshold and search quit threshold and propose it should be adapted to explain these observed patterns of accuracy with time on task. What is typically thought of as standard and robust research findings in controlled laboratory settings may not directly apply to real-world environments and instead large, controlled studies in relevant environments are needed.

Significance

For over 70 years, researchers have believed that as time on search tasks increase humans make more errors detecting target 'events' (and take longer): a 'vigilance decrement'. Previous research on this has been undertaken in laboratory settings, on tasks with little control over presentation rate, but generalized to real-world scenarios, leading to regulations limiting continuous viewing time in cancer screening. We demonstrate in a large, controlled study in clinical practice, where readers self-pace reading and rest breaks, reduced accuracy is not observed. Overall accuracy increases with time on task with fewer false alarms. Instead of limiting continuous viewing time, work environments for breast screening should allow experts uninterrupted sessions of self-chosen length, thus improving accuracy and reducing unnecessary further tests.

Main Text:

Errors in search and monitoring tasks have devastating consequences. In an undercover operation in airport baggage screening, operators failed to detect over 70% of mock knives, guns and explosives(1). Expert radiologists make an average of forty million errors interpreting medical images worldwide annually(2),(3). Errors by expert radiologists substantially contribute to diagnostic error(2), which causes 80,000 deaths per year in the US(4). Across Europe,

Australasia and North America 90,000 women each year have breast cancer missed by expert radiologists during mammography examination (0.07% to 0.15%(5-7) of >60million women screened(8-12)) and up to 7 million women have unnecessary further tests after false positive mammography decisions (6.5% to 12.1%)(7, 13, 14). The mechanisms of these miss errors are multiple and complex. One of the earliest and most studied proposals was an increase in miss errors over time on task, called the 'vigilance decrement'(15).

The first evidence of a vigilance decrement was in radar operators in World War II, where detection of aircraft and submarines dropped after 30 – 45 minutes of their shift(16). This led to the seminal studies by Mackworth where RAF observers monitored a specialised clock hand for a 'signal' (the hand jumping forward two positions, rather than one) over a period of 2 hours(15, 17, 18). Observers again showed a drop in performance after 30 minutes, prompting the authors to recommend that shifts should be limited to 30 minutes for tasks that require constant vigilance. A large body of literature from the field of psychology has used similar abstract tasks to determine the circumstances under which vigilance decrements are of greatest magnitude (i.e. when event rates are high, and/or successive discriminations are required(19)). Explanations are broadly based around cognitive overload and underload(19, 20). The vigilance decrement has been observed in a range of experiments that approximate real-world activities, e.g. airport baggage screening(21), assembly line inspection(22), driving(23), radar operation(24), and interpretation of medical images(25). However, these studies lack ecological validity because participants know that they are taking part in research so the jeopardy of missing a cancer or allowing dangerous items onto an aeroplane is absent. Indeed, there is evidence from medical imaging that performance during experimental studies is not reflective of performance in clinical practice (the laboratory effect)(26-28). Despite the limitations of current research, procedures aimed at reducing vigilance decrements have been developed and implemented(29-31). These include regulations for regular breaks in cervical screening (breaks every 10 – 15 minutes)(30) and limiting continuous viewing in airport baggage security (breaks every 20 – 30 minutes)(31).

Appropriate health and safety guidance for safety critical monitoring tasks relies on knowing whether the vigilance decrement contributes to miss errors, or is simply a product of the research methods used in previous studies(32). This requires evidence from real-world studies where laboratory-effect biases are not present. Such research is challenging, because these tasks often

involve searching for rare targets, which require very large studies with designs that do not interfere with safety critical tasks. Despite a wealth of research on vigilance decrement (608 studies indexed in Medline), only four real-world studies have investigated this, two in radar surveillance(16, 33), one in baggage security (34) and one in breast cancer screening (35). In the first radar surveillance study, a brief report with little details of the methods, and the number of participants not reported, vigilance decrement was observed at 30 minutes (for radar operators searching for submarines) and 45 minutes (for radar operators searching for aircraft)(16). In the second radar surveillance study, no vigilance decrement was observed amongst 16 radar operators, in which simulated data were mixed into live air traffic data(33). In the baggage security study, x-ray screeners were divided into two groups: one who screened for 20 minutes (i.e., their usual working conditions) and one who screened for up to 60 minutes but could decide to take a break(34). Simulated threats were added to the images., which is standard practice in airport security. No difference was observed in the percentage of correctly detected images of simulated threat items between the two groups. In the breast cancer screening study, observational data on the interpretation of mammograms from 610,104 women read by 148 radiologists were extracted from the Norwegian Breast Cancer Screening programme (35). The effect on reader performance of the position of an image (ranging from 10^{th} image to 300^{th} image) within a batch was assessed, where a batch was defined as a reading sequence that lasted until a break of 15-minutes or more occurred between two interpretation decisions. There was some evidence for a vigilance decrement shown as a small but statistically significant reduction in true positive (cancer detected) interpretations with time on task (0.2/1000 decrease over first 100 cases, 5% relative to first 10 cases), but this study excluded missed (interval) cancer so could not assess test sensitivity. There was also a concurrent larger decrease in false positive recalls (10/1000 decrease over first 100 cases, 19% relative to first 10 cases). However, these are purely observational data with no analysis of whether this is driven by a radiologist vigilance decrement or confounding factors such as radiologists moving more difficult cases for later consideration or women who are more likely to have cancer being allocated to the first half of batches.

A large randomised controlled trial of >1 million women attending the English mammography screening programme investigated an intervention which changed the order in which cases were examined to reduce the impact of the vigilance decrement(36). The intervention had no effect in short work sessions(37). These data have the advantage of bespoke trial software to detect and

correct for potential confounding such as radiologists moving cases, and the intervention reversed case order enabling analysis of potential systematic biases in risk of cancer with batch position. In the current work we use data from that study to determine whether experts examining breast screening mammograms for signs of cancer experience a vigilance decrement, in both short and long work sessions.

Results & Discussion:

We studied breast screening mammograms from 1,069,566 women (mean age 59 years), of whom 226,506 (21%) were attending their first ever screening appointment. Each woman's mammograms were independently examined for signs of cancer by two qualified specialist experts (henceforth referred to as 'experts'). In all, 360 experts are included in this study. 8,761 (0.82%) of the women had cancer detected at screening and a further 2,046 (0.21%) had cancer detected symptomatically within 3 years of screening. Further descriptive statistics appear in SI Appendix (table S1 and figure S1).

Mammography speed and accuracy improves with time on task

The vigilance decrement predicts a reduction in cancer detection rate (number of cancers detected by the expert per thousand women screened) with time on task. We found that neither cancer detection rate (figure 1a-b) or test sensitivity, (proportion of women with cancer who were detected by screening, figure 2a) changed over the course of examining 200 women's mammograms since their last break of 20 minutes or more(Odds Ratio (OR) [5 extra women's mammograms since the expert's last break]= 0.998 (95%CI 0.994-1.0008)). This pattern was observed for our main definition of a break (20 minutes or more without inputting a decision into the computer, colored orange in figures 1 and 2) and our sensitivity analyses defining a break as >10, >60, >180, or >480 minutes without inputting a decision (black, blue green and pink inon figures 1 and 2, respectively). Examining each woman's mammograms takes a median 36 seconds (mean 69 seconds, distribution in the SI Appendix, figure S5), so this translates to no vigilance decrement observed after more than 2 hours on task for most experts, and over one hour for the faster experts.

Working for more than an hour without a break resulted in improvements in overall accuracy rather than decrements. This was due to a reduction in false positive recalls (false alarms where a woman without cancer is incorrectly recalled for further tests, which causes her anxiety and

consumes significant resources). The overall recall rate (proportion of women recalled for further tests, using the 20 minute break definition in figure 1c-d) decreased rapidly from 4.66%, (95% CI 4.23% - 5.12%) when the expert started the task, to 3.99% (3.63% - 4.39%) when examining the 40th woman's mammograms, to 3.69% (3.35% - 4.06%) when examining the 100th woman's mammograms and 3.24% (2.90% - 3.62%) when examining the 200th woman's mammograms without taking a break 20 minutes or more. This is clinically and operationally significant, as in a national programme screening 2 million women/year the difference between a recall rate of 3.24% and 4.66% is an additional 28,400 (44%) unnecessary false positive recalls to assessment. This was similarly reflected in increasing test specificity (proportion of women without cancer who are correctly told they do not have cancer, figure 2b), and in increasing positive predictive value (the proportion of women recalled for further tests who have cancer, figure 2c) with time on task.

In addition to becoming more accurate, experts also made each decision more quickly as the number of women's mammograms examined since their last break increased (a measure of time on task, figure 1e-f and SI Appendix, table S3). At the start, experts took a mean of 73.7s (95%CI 73.4s - 73.9s) to examine each woman's mammograms. By the 20th woman's mammograms examined since their last break of at least 20 minutes, this had reduced to 64.4s (95%CI 64.1s - 64.6s), to 60.6s (95%CI 60.4s – 60.9s) by the 100th woman's mammograms and to 55.1s (95%CI 54.9s - 55.4s) by the 200th woman's mammograms. Full model results are in the SI Appendix (tables S2 to S5).

Changes in accuracy are dependent on break length

After longer breaks the experts start the session with a higher recall rate than after shorter breaks, and they decrease at a similar rate. When reading the first session of a working day following at least 8 hours without reading activity (the >480-minute-break definition, fig 1c-d) the recall rate is initially high (5%,) and reduces with time on task (3.5% at the 200th woman). When including short breaks of 20 minutes or more, the expert has partially reset their recall rate to be higher again (4.7%, fig 1c-d, >20minute definition), and then it declines again with time on task (3.2% at the 200th woman). Similarly, specificity is lower at the start of a working day (95.3%, fig 2b, >480 definition, group 1: first 30 cases) and increases with time on task (96.5%, fig 2b group 4: cases 91 to 200). After a 20-minute break or longer, specificity is reduced but not as much as at

the beginning of the day (95.7%, fig 2b, group 1: first 30 cases). Time taken per case follows the same pattern of being highest at the beginning of a working day, decreasing with time on task, with breaks within a day not fully resetting to match time taken at the beginning of the day (fig 1e-f). This is further explored in the SI Appendix (figure S4), which shows the same patterns when breaks of <1hour, 1-3 hours, 3-12 hours and >12 hours are analyzed separately.

Patterns of increasing accuracy are robust when considering bias and statistical power

These are observational data, so consideration must be given to whether these effects might be driven by measured or unmeasured confounders or biases. First, were women whose mammograms were examined first in a reading session systematically different to those examined later? Whilst the allocation system suggests no reason for this, in large well powered datasets it is important to examine this empirically. This was tested using data from the intervention arm of the original trial, where up to 111 women's mammograms were grouped together in 'sessions'. The first and second experts examined each session in the opposite order to one another, yet the reduction in recall rate and time taken were observed with both experts (see figure 3b and 3c). Therefore, the effects are unlikely to be due to confounding associated with the woman's characteristics.

Confounding at the expert level was also considered, as experts themselves could choose the length of time on task since their last break. Were experts who chose to take more breaks systematically different from those who took fewer? Specifically, were experts who read long sessions quicker and have lower recall rate, and it was the amalgamation of different session lengths giving the appearance of decrease over time? Analyses were repeated for each session length, for example only including task sessions when a given minimum number of mammograms were examined. The same effect of decreasing recall rate and decreasing time taken to examine each women's mammograms was found for every session length (figure 3d and 3e and the SI Appendix, figure S2). This demonstrates that these effects are not caused by expert level confounding, because they were still present when fixing session length. Further we investigated whether effects were caused by experts changing the order in which they examined the women's mammograms, and this was also not causing confounding as shown in SI Appendix (figure S3).

Theoretical explanations for changes with time on task

Signal detection theory is the key framework underpinning most models for understanding performance and error in medical imaging. Within this framework, the improvement in accuracy with time on task observed here could be due to the experts' fundamental accuracy increasing, characterized by moving to a higher Receiver Operating Characteristic (ROC) curve (an improvement in ability to discriminate between mammograms with and without cancer), or due to a change in decision threshold (becoming more reluctant to recall cases) within the same fundamental ROC curve. Whilst overall accuracy does improve with time on task this still may be explained by a change in decision threshold if the experts are becoming more reluctant to recall whilst operating on a low gradient part of the ROC curve (i.e., they are making very inclusive decisions to recall many women, so changing their decision threshold only excludes women with minimal signs of cancer). A previous study in the Norwegian breast screening programme did find a small reduction in true positive recalls with time on task, alongside a larger reduction in false positive recalls(35). Their programme is similar to the UK's, but overall cancer detection rate in their study was 42 per thousand women screened, whereas in our study it was 88 per thousand women screened. It is possible that both cohorts experienced a threshold shift with time on task, but in Norway this affected cancer detection to a greater extent because they are operating at a different point on the ROC curve where a threshold shift will have a greater effect on cancer detection rate due to a differing slope. However, this comparison is confounded by the different population risk profiles and screening intervals.

The Multiple-Decision Model (MDM) of search(38) (see Figure 4), uses signal detection theory to explain the impact of disease prevalence on accuracy. This model proposes that at low prevalence (as found in cancer screening), experts do not change their fundamental accuracy (as measured by d-prime), but they spend less time searching the display and have increased decision thresholds (so are more willing to say that a mammogram is cancer-free)(39). We propose applying the Multiple Decision Model to time on task in a similar manner to low prevalence. This would predict a universal reduction in reading time due to a lowering of the 'quitting threshold' for search over time (i.e., experts will spend less time searching each display before making a decision), as indicated in figure 1e and 1f. It also predicts a change in decision threshold over time so experts would be less willing to say there is a cancer, without a corresponding change in d-prime – consistent with our signal detection data shown in Figure 2e

and f and the increase in specificity over time as seen in Figure 2b. The Criterion (decision threshold) in figure 2f may also be higher at the start of sessions preceded by a shorter break, which might indicate short breaks do not result in a full reset of decision threshold or reduction in specificity compared to long breaks. This is explored in the SI Appendix (figure S4) which shows that specificity and criterion are lower after a long break of >12hours, compared to after a short break of <1hour. The mammography task is thought to consist of three stages, search (global processing then targeted search to locate regions of interest), recognition, and decision(40). There is evidence that experts can extract important global information from an image within the first glance to determine whether an image contains an anomaly or not (in some instances within the first 250ms of an image being presented)(41, 42). This 'gist' of information is likely to shape an expert's subsequent search so that they are more effective at detecting cancer and less likely to introduce additional false positive recalls, through focusing on the most important areas of the mammogram first. In these circumstances the effect of shortening of search on the accuracy of the highly experienced experts in this study is reduced. In the Norwegian study the speed of reading increased with time on task in a similar way to our results, and their modelling indicated that 17% of the reduction in true positive results was mediated by reading speed. In the UK, time spent reading each case was longer than in Norway, which may explain why the increased reading speed did not have the same impact. Whilst this explanation is a good fit to these data, it doesn't preclude other explanations. It does suggest that considerations of break scheduling are more complex than a simple reduction in vigilance and should also consider threshold shifts from the current threshold and related clinical outcomes.

Initial data on radar monitoring(16), and successive decades of research, show a clinically and statistically significant decrement in detection with time on task. Why do we see a different pattern in screening mammography, predominantly of improved performance? Within mammography, search for an anomaly is perceptually more varied and complex, compared to radar surveillance (and similar lab-based experiments), but perhaps most importantly signals are not time-limited or transient in nature. Furthermore, experts may be intrinsically motivated in their goal to reduce illness (as opposed to being told by superiors/experimenters to find a target). Both of these factors are known to reduce the vigilance decrement with some suggestion that the vigilance decrement may be a byproduct of the vigilance study design (20, 32). Radiologists may also choose to look at a mammogram for as long as needed (within reason to ensure they still

read their large volume of images within a batch, see the SI Appendix (figure S5) for distribution of time taken per case). In contrast, tasks like radar monitoring and the original vigilance studies used stimuli where the timing of the target was controlled by the experiment rather than the reader (e.g. Mackworth, 1944, (18)). This may have an effect on the vigilance decrement. Indeed, a study involving 22 baggage screeners examining images with simulated threats at an international airport, where the timing was controlled by the expert, showed limited evidence for vigilance decrements when the task load (rate of presentation of baggage images) was low to medium but this developed as task load became high (significant interaction between task load and time on task)(34)).

Similarly, examining mammograms is time pressured due to reading volume overall but experts can determine their own break and task scheduling, and there is evidence that breaks (either enforced or self-selected) can reduce the vigilance decrement. In the laboratory, the introduction of breaks has been shown to reduce the vigilance deficit for both self-paced breaks(43) and those that are imposed (e.g., Helton & Russell, 2015(44); Ross, Russell & Helton, 2014(45), though this is not a universal finding(46)). For example, Helton and Wen ((47)) suggest that the addition of a break can lead to a renewal of resources that would otherwise be depleted according to resource depletion theories of the vigilance decrement (see also (48)). As experts can regulate rest breaks more than would typically be expected in laboratory vigilance studies, this could be a factor in why we saw no vigilance decrement in our mammogram data. In fact, readers in an X-ray baggage screening task reported more engagement in the task when they could choose for themselves when to stop, suggesting that control of when to take breaks is in itself beneficial(34).

The practical purpose of understanding the underlying mechanisms is so we can generalize to other medical imaging tasks, and broader search tasks. These tasks will vary in disease prevalence, recall threshold, accuracy, expertise of experts, and reading environment. The current findings demonstrate that vigilance decrement theory should not be applied without empirical evidence in the real-world setting. Similarly, we cannot assume that the accuracy and speed improvements with time on task reported here are generalizable to other tasks requiring vigilance, though they are most likely to be generalizable to self-paced radiology and pathology tasks.

Conclusions

We found no significant vigilance decrement experienced by qualified experts examining up to 200 women's mammograms sequentially. Instead, performance improves with time on task through increased speed and a reduction in number of false positive recalls for assessment. Further, shorter breaks were associated with fewer false positive recalls to assessment at the beginning of the next batch compared to longer breaks Our results suggest that population breast screening programmes should enable experts to review several sessions of up to 200 women's mammograms consecutively if they wish, with self-selected break scheduling, minimize interruptions in the work environment. This contrasts, in part, with the results of the only other study on this topic which found a small but statistically significant decrease in sensitivity with time on task(35). Both studies found an increase in specificity and decrease in time taken per case with time on task. In combination both of these studies support a Multiple Decision Model with a threshold shift to explain behavior changes with time on task, rather than a vigilance decrement.

Determining the extent to which our findings generalize to other repetitive clinical tasks such as those within other screening programmes and real-world vigilance tasks will require further applied research. However, this study demonstrates that a huge literature of laboratory-based psychological research incorrectly predicts what may happen in at least one clinical practice task – that of mammography reading. Further research is required to understand the underlying psychological mechanisms, and how they impact real-world health outcomes. Laboratory and real-world research should be brought together through interdisciplinary collaborations to answer these questions.

Materials and Methods

Methods

This observational study uses data from a randomised controlled trial; Changing case Order to Optimise patterns of Performance in Screening (CO-OPS, ISRCTN46603370), which is reported in detail elsewhere(37). Ethical approval was granted by the Coventry and Warwickshire National Health Service (NHS) Research Ethics Committee on June 27, 2012 (ref WM/0182). Approval to archive the data and carry out further analysis on the dataset was granted on June 7th 2022 (ref 12/WM/018). Informed consent was at the centre level by the director of breast

screening, as the intervention and control group were considered both to be standard practice. Women's mammograms were examined in sessions (median size 35 women) grouped together within the computer software, with two experts independently examining each session in either the opposite order to one another (intervention) or the same order (control). In practice many experts examined several sessions sequentially without a break. In this study we examined accuracy with number of women's mammograms examined since the expert's last break (as a proxy for time on task). We defined a break as either 10 minutes, 20 minutes, 60 minutes, 180 minutes and 480 minutes without inputting a decision into the computer software. The primary definition was 20 minutes. The outcomes of recall rate, cancer detection rate and time taken to read were modelled to assess the effect of number of women's mammograms examined since a break; using screening centre and expert as levels in a multilevel model.

Study population

The CO-OPS trial involved 46 breast cancer screening centres in England, UK, for one year commencing in 2012. The practice in those centres was for two experts to independently examine women's mammograms to decide whether to recall her for further tests. Discordant decisions were resolved by arbitration, and in some cases decisions by both experts to recall were also arbitrated. Most centres did not blind the second expert to the decision of the first. Experts were all radiologists, breast clinicians or radiography advanced practitioners or consultants qualified to read in the NHS breast screening programme(49), requiring a minimum of 5000 women's mammograms to be examined per year. Within the computer software mammograms from all women screened at a single location on the same day were presented together as a list, we refer to this as a session. The expectation was that experts would examine a whole session without a break. In practice many experts report examining several sessions sequentially without a break. The CO-OPS trial questioned whether the second expert examining the session in the reverse order to the first expert would change the cancer detection rate, via vigilance decrements occurring for each expert at different points in the session i.e., whilst examining different women's mammograms. The intervention was not effective(37). The observational analysis reported here examines the patterns of performance with time on task, using the time stamp from the computer software of each expert's decision to recall or not. Every woman screened by each centre as part of the NHS Breast Screening Programme during the trial period was included in

the trial and this subsequent observational analysis; with 1,194,147 women involved in this final analysis.

Outcomes

This study uses three main outcome variables: the proportion of women that the expert recommended recalling for further tests (recall rate), the proportion of women in which the expert detected a cancer (cancer detection rate, requiring both the expert to suggest recall and the follow-up tests indicate biopsy proven breast cancer) and the time taken for the expert to examine and decide whether to recall the woman. Outcome data for these three variables are complete: recall decisions and a time stamp for when they were made were automatically populated in the software and form part of workflow at the centre, and cancers detected after recall from screening and proven by biopsy will be complete due to the standardized quality assurance processes. We explore the effect that time on task (characterized as number of women's mammograms examined since a break) has on these three outcomes. We also examine test accuracy using the reference standard of biopsy-proven breast cancer either after recall from the screening appointment, or symptomatically detected within 3 years after screening. Symptomatically detected cancers were communicated back to screening centres from the English cancer registry. Women who moved abroad, had cancer detected after longer than 3 years, or in whom the cancer registry failed to report back to the screening centre would not have been included in this dataset. Women who did not have cancer detected at screening and did not have a record of a symptomatic cancer in the 3 years after screening were assumed to not have cancer. Screening test outcomes are defined as true positive (TP, the woman had cancer which was detected), false negative (FN, the woman had cancer which was missed), false positive (FP, the woman did not have cancer, she was worried unnecessarily after incorrect recall for further tests) and true negative (the woman did not have cancer and was correctly reassured). The standard four test accuracy metrics were calculated as follows: sensitivity = TP/(TP+FN), specificity = TN/(TN+FP), positive predictive value (PPV) = TP/(TP+FP), and negative predictive value (NPV) = TN/(TN+FN).

Data preparation

The trial was implemented through adaptations to the UK National Breast Screening Service's (NBSS) computer system, which records each expert's identity and decision for every case alongside any subsequent arbitration decision and whether cancer was detected following recall from screening. The system also records whether cancer was detected symptomatically in the years following screening, but this requires human input through a national system based on cancer registry data, so may be incomplete. The NBSS computer system was adapted for the CO-OPS trial to record additional variables, including date and time of each decision, and to detect when cases were not read within the intended order within a session.

Experts often examine more than one session in succession, and performance over time periods longer than a single session is of interest. Therefore, we used the exact time and date of each decision to combine sessions examined subsequently by the same expert into a single session, using different assumptions regarding the time period that had elapsed without a decision (and therefore assumed to represent the reader taking a break). We considered the primary definition of a break as 20 minutes without a decision as experts would very rarely take longer than 20 minutes to make and report a decision. We undertook sensitivity analyses by additional shorter or longer break time definitions (10, 60, 180 or 480 minutes) to check the robustness of this assumption. Unless otherwise stated results are given for the 20-minute definition.
We used cases examined as the first or the second expert to establish the case order, but only used decisions made as the first expert in our analysis to preserve independence of cases in the models and to ensure decisions were made independently without consulting the other expert's decision. We excluded cases that were not examined in the intended order. We undertook a sensitivity analysis to check whether including these moved cases changed conclusions and it did not (see SI Appendix, figure S3). Data preparation is described in more detail in the methods section of the SI Appendix.

Statistical Analysis

Multilevel models were used; with the individual mammogram as level one, expert as level two (to account for the effect of different individual experts) and centre as level three. Generalised linear models were used; with recall and cancer detection (binary outcomes) analysed with

logistic models and time taken on the case analysed with a gamma distribution. The main explanatory variable studied was position in session, as defined above (as a proxy for time on task), with models evaluated for each definition of session position, determined by the different break definitions. Position in session was included in the models for recall and time taken to read using a linear basis spline, with knot points at positions 20 and 40, but as a linear term for cancer detected, because the pattern for cancer detected was not curved in shape. The age of the woman at the time of the mammogram and whether it was the woman's first mammogram (prevalent) or a subsequent one (incident) were also used as explanatory variables.

Each model was run on a subset of the dataset, removing all mammograms read either first in the session (as these may be systematically different based on our session definition) or after position 200 in the session (as these represented influential outliers). The number of mammograms removed for these reasons is different for each break time definition and are shown in the SI Appendix (table S1). For the models of time taken to read, mammograms for which the time taken to read was greater than 10 minutes (and some cases where the time was recorded as zero) were also removed from the dataset.

To examine the possibility that there was some confounding present from systematic differences in characteristics of women examined early and late in the session we undertook two extra analyses. Firstly, we analysed pattern of performance over the course of the session comparing the cases from the intervention trial arm, those examined in forward and those examined in reverse order by reader 1. These were sessions examined in the opposite order, so thus we separated the effect of any confounder associated with the characteristics of the women screened at different session positions with the effect of time on task. The analysis was performed using R software, with the multilevel models being fitted with the "lme4" package. Statistical significance was assessed at the 5% level.

Confidence intervals for sensitivity, specificity, positive predictive value and negative predictive value were calculated using the Wilson method for binomial confidence intervals.
The measure d' (d prime) is calculated by $d' = z(Sensitivity) - z(1 - Specificity)$, and criterion by $c = -\frac{1}{2}(z(Sensitivity) + z(1 - Specificity))$, where z is the inverse cumulative

normal distribution function, as given in Macmillan and Creelman(50). The confidence intervals were calculated using the approximation given by Gourevitch and Galanter (1967)(51).

Acknowledgements: We are extremely grateful for the hard work of all of the clinical and administrative staff at the screening centres who participated in the original CO-OPS trial.

Funding: The CO-OPs RCT is independent research arising from an NIHR Postdoctoral fellowship and an NIHR Career Development Fellowship for STP (CDF-2016-09-018). The trial is registered under ISRCTN46603370 and received Research Ethics Committee approvals (ref WM/0182). The study sponsor is the University of Warwick. STP, KF and CS disclose additional support for the publication of this work from the NIHR (Research Professorship, NIHR302434 [STP, KF, CS] and Development and Skills Enhancement Fellowship, NIHR302371 [KF]). This research was supported by the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014 [MW]). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Data and materials availability:

These data that support the findings of this study are available on reasonable request from the corresponding author [STP], conditional upon that request meeting the requirements of the study ethical approvals. These data are not publicly available due to the inclusion of patient data.

Code availability: The code are available upon reasonable request from the corresponding author [STP].

References

1. CBS News (2017) TSA screenings fail to spot weapons most of the time, agency says.
2. M. A. Bruno, E. A. Walker, H. H. Abujudeh, Understanding and Confronting Our Mistakes: The Epidemiology of Error in Radiology and Strategies for Error Reduction. *Radiographics* **35**, 1668-1676 (2015).
3. L. Berlin, Accuracy of diagnostic procedures: has it improved over the past five decades? *AJR Am J Roentgenol* **188**, 1173-1178 (2007).
4. M. L. Graber, The incidence of diagnostic error in medicine. *BMJ Qual Saf* **22 Suppl 2**, ii21-ii27 (2013).
5. P. C. Allgood, S. W. Duffy, R. Warren, G. Hunnam, Audit of negative assessments in a breast-screening programme in women who later develop breast cancer-implications for survival. *Breast* **15**, 503-509 (2006).

6. M. A. Durand *et al.*, False-Negative Rates of Breast Cancer Screening with and without Digital Breast Tomosynthesis. *Radiology* **298**, 296-305 (2021).
7. H. D. Nelson, E. S. O'Meara, K. Kerlikowske, S. Balch, D. Miglioretti, Factors Associated With Rates of False-Positive and False-Negative Results From Digital Mammography Screening: An Analysis of Registry Data. *Ann Intern Med* **164**, 226-235 (2016).
8. Australian Institute of Health and Welfare (2019) BreastScreen Australia monitoring report 2019.
9. Canadian Partnership Against Cancer (2017) Breast Cancer Screening in Canada: Monitoring & Evaluation of Quality Indicators.
10. International Agency for Research on Cancer (2017) Cancer Screening in Report on the implementation of the Council Recommendation on cancer screening.
11. New Zealand National Screening Unit, Breast Screening DHB quarterly reports. (2021).
12. United States Food and Drug Administration (2019) MQSA National Statistics
13. R. L. Bennett, S. J. Sellars, S. M. Moss, Interval cancers in the NHS breast cancer screening programme in England, Wales and Northern Ireland. *Br J Cancer* **104**, 571-577 (2011).
14. T. H. Ho *et al.*, Cumulative Probability of False-Positive Results After 10 Years of Screening With Digital Breast Tomosynthesis vs Digital Mammography. *JAMA Netw Open* **5**, e222440 (2022).
15. N. H. Mackworth, The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology* **1**, 6-21 (1948).
16. Anonymous. (1944) Changes in efficiency during ASV watches. (RAF/ORS/CC).
17. N. H. Mackworth, Researches on the measurement of human performance. *Researches on the Measurement of Human Performance.* (1950).
18. N. H. Mackworth (1944) Notes on the clock testnew approach to the study of prolonged visual perception to find the optimum length of watch for radar operator. (Medical Research Council Report).
19. J. E. See, S. R. Howe, J. S. Warm, W. N. Dember, Meta-analysis of the sensitivity decrement in vigilance. *Psychological Bulletin* **117**, 230 (1995).
20. D. R. Thomson, D. Smilek, D. Besner, Reducing the vigilance decrement: The effects of perceptual variability. *Conscious Cogn* **33**, 386-397 (2015).
21. K. M. Ghylin, C. Drury, R. Batta, L. Lin (2007) Temporal effects in a security inspection task: Breakdown of performance components. in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (SAGE Publications Sage CA: Los Angeles, CA), pp 93-97.
22. R. V. Badalamente, M. M. Ayoub, A behavioral analysis of an assembly line inspection task. *Hum Factors* **11**, 339-352 (1969).
23. J. C. Verster, T. Roth, Vigilance decrement during the on-the-road driving tests: The importance of time-on-task in psychopharmacological research. *Accident Analysis & Prevention* **58**, 244-248 (2013).
24. D. Lindsley, I. Anderson, Radar operator'fatigue': The effects of length and repetition of operating periods on efficiency of performance. *Office of Scientific Research Research and Development*, 566-572 (1944).
25. S. Taylor-Phillips *et al.*, Retrospective review of the drop in observer detection performance over time in lesion-enriched experimental studies. *Journal of digital imaging* **28**, 32-40 (2015).
26. D. Gur *et al.*, The “laboratory” effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology* **249**, 47-53 (2008).
27. D. L. Miglioretti *et al.*, Correlation between screening mammography interpretive performance on a test set and performance in clinical practice. *Academic radiology* **24**, 1256-1264 (2017).
28. C. M. Rutter, S. Taplin, Assessing mammographers' accuracy: a comparison of clinical and test performance. *Journal of clinical epidemiology* **53**, 443-450 (2000).

29. D. Laming, R. Warren, Improving the detection of cancer in the screening of mammograms. *Journal of Medical Screening* **7**, 24-30 (2000).
30. National Health Service (2003) Laboratory Organisation: A Guide for Laboratories Participating in the NHS Cervical Screening Programme NHSCSP Publication No 14.
31. European Commission (2015) Commission implementing regulation (EU) 2015/1998 of 5 November 2015 laying down detailed measures for the implementation of the common basic standards on aviation security.
32. P. A. Hancock, In search of vigilance: the problem of iatrogenically created psychological phenomena. *Am Psychol* **68**, 97-109 (2013).
33. R. A. Pigeau, R. Angus, P. O'Neill, I. Mack, Vigilance latencies to aircraft detection among NORAD surveillance operators. *Human Factors* **37**, 622-634 (1995).
34. D. Buser, A. Schwaninger, J. Sauer, Y. Sterchi, Time on task and task load in visual inspection: A four-month field study with X-ray baggage screeners. *Applied Ergonomics* **111**, 103995 (2023).
35. H. A. Backmann, M. Larsen, A. S. Danielsen, S. Hofvind, Does it matter for the radiologists' performance whether they read short or long batches in organized mammographic screening? *Eur Radiol* **31**, 9548-9555 (2021).
36. S. Taylor-Phillips *et al.*, Changing case Order to Optimise patterns of Performance in mammography Screening (CO-OPS): study protocol for a randomized controlled trial. *Trials* **15**, 17 (2014).
37. S. Taylor-Phillips *et al.*, Effect of using the same vs different order for second readings of screening mammograms on rates of breast cancer detection: a randomized clinical trial. *Jama* **315**, 1956-1965 (2016).
38. J. M. Wolfe, M. J. Van Wert, Varying target prevalence reveals two dissociable decision criteria in visual search. *Curr Biol* **20**, 121-124 (2010).
39. J. M. Wolfe *et al.*, Low target prevalence is a stubborn source of errors in visual search tasks. *J Exp Psychol Gen* **136**, 623-638 (2007).
40. H. L. Kundel, C. F. Nodine, D. Carmody, Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Invest Radiol* **13**, 175-181 (1978).
41. K. K. Evans, D. Georgian-Smith, R. Tambouret, R. L. Birdwell, J. M. Wolfe, The gist of the abnormal: above-chance medical decision making in the blink of an eye. *Psychon Bull Rev* **20**, 1170-1175 (2013).
42. H. L. Kundel, C. F. Nodine, Interpreting chest radiographs without visual search. *Radiology* **116**, 527-532 (1975).
43. A. Chavaillaz, A. Schwaninger, S. Michel, J. Sauer, Work design for airport security officers: Effects of rest break schedules and adaptable automation. *Applied Ergonomics* **79**, 66-75 (2019).
44. W. S. Helton, P. N. Russell, Rest is best: the role of rest and task interruptions on vigilance. *Cognition* **134**, 165-173 (2015).
45. H. A. Ross, P. N. Russell, W. S. Helton, Effects of breaks and goal switches on the vigilance decrement. *Experimental Brain Research* **232**, 1729-1737 (2014).
46. G. E. Waldfogle, A. E. Garibaldi, A. R. Neigel, J. L. Szalma, 'I need a break': the effect of choice of rest break duration on vigilance. *Ergonomics* **64**, 1509-1521 (2021).
47. W. S. Helton, J. Wen, Will the real resource theory please stand up! Vigilance is a renewable resource and should be modeled as such. *Experimental Brain Research* **241**, 1263-1270 (2023).
48. J. S. Warm, R. Parasuraman, G. Matthews, Vigilance Requires Hard Mental Work and Is Stressful. *Human Factors* **50**, 433-441 (2008).
49. NHSBSP (2011) Quality assurance guidelines for breast cancer screening radiology. (NHSBSP.).
50. N. A. Macmillan, C. D. Creelman, *Detection theory: A user's guide* (Psychology press, 2004).
51. V. Gourevitch, E. Galanter, A significance test for one parameter isosensitivity functions. *Psychometrika* **32**, 25-33 (1967).

Figures:

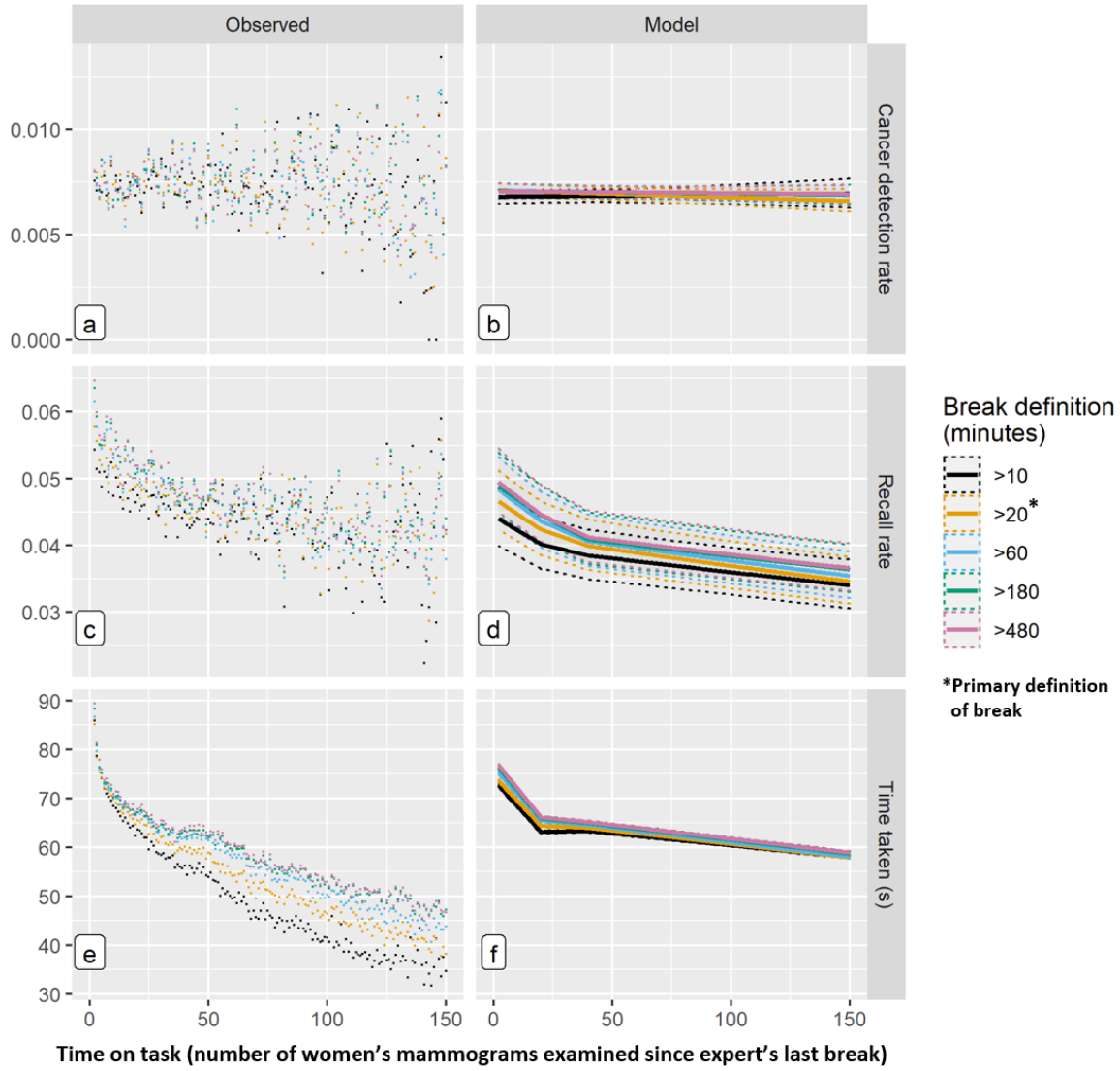

Figure 1. Performance metrics with time on task, represented by the number of women's mammograms examined consecutively since the expert's last break. Observed values are shown to the left and multilevel model fitted values (shown as a solid line with 95% confidence intervals as dotted lines) to the right. Models were adjusted for women's age and whether they have previously attended screening, with clustering for expert and screening center. The primary definition of a break was 20 minutes without inputting a decision on the computer. Sensitivity analyses exploring different definitions of what constituted an expert's break: 10, 60, 180, or 480 minutes are shown in different colors. (1a) cancer detection rate data calculated as the proportion of women that were correctly identified as having cancer (1b) modelled result; cancer detection rate; (1c) recall rate data (calculated as the proportion of women the expert indicated required recall for further tests) (1d) modelled recall rate result; (1e) mean time taken to examine each woman's mammograms data and (1f) modelled mean time. For example, an orange data point with x=50 includes data from every woman who was examined at the 50th position in the session after a break of 20 minutes or more. If the definition of a break is changed to at least 60 minutes, shorter breaks are ignored, and some women are re-categorised into larger sessions for analysis (blue). Because of this re-categorisation for sensitivity analyses of different break definitions, women's mammograms can appear more than once as represented in the different coloured analyses but only once in each color.

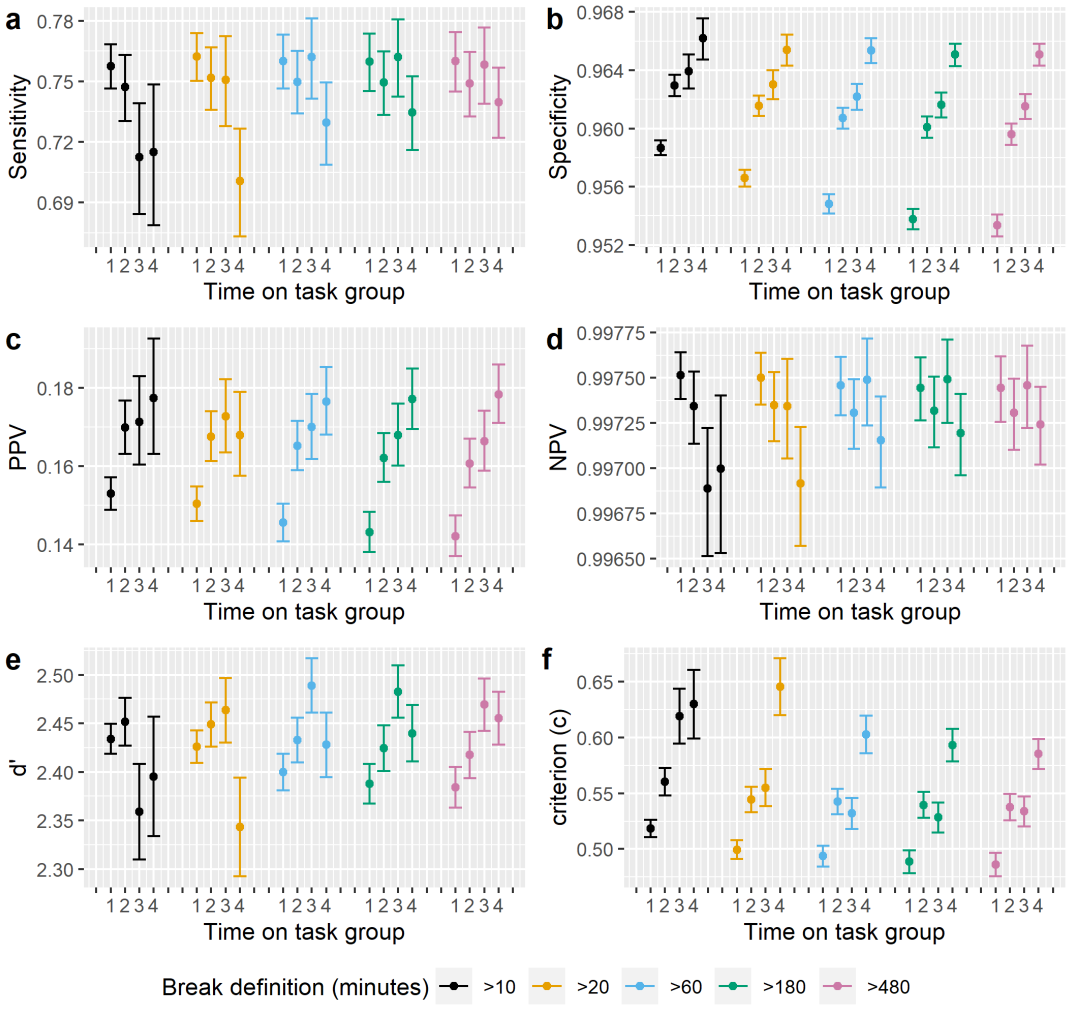

Figure 2. Test accuracy represented by a. sensitivity, b. specificity, c. positive predictive value (PPV) d. negative predictive value (NPV), e. d-prime f. criterion (decision threshold/willingness to recall women for further tests), with time on task. Bars represent 95% confidence intervals. Data in orange represent the primary analysis using a break definition of 20 minutes. Time on task is represented by how many women's mammograms have been examined with a 20-minute break. Data were analyzed in 4 groups based on the position number of mammograms in a session: group 1 – mammograms that were examined at the 2nd to 30th position in the session after a break of at least 20 minutes, group 2 – mammograms that were examined at the 31st to 60th position in the session after a break of at least 20 minutes, group 3 – mammograms that were examined at the 61st to 90th position in the session after a break of at least 20 minutes, group 4 – mammograms that were examined at the 91st to 200th position in the session after a break of at least 20 minutes. Sensitivity analyses exploring alternative definitions of break length are shown in different colors.

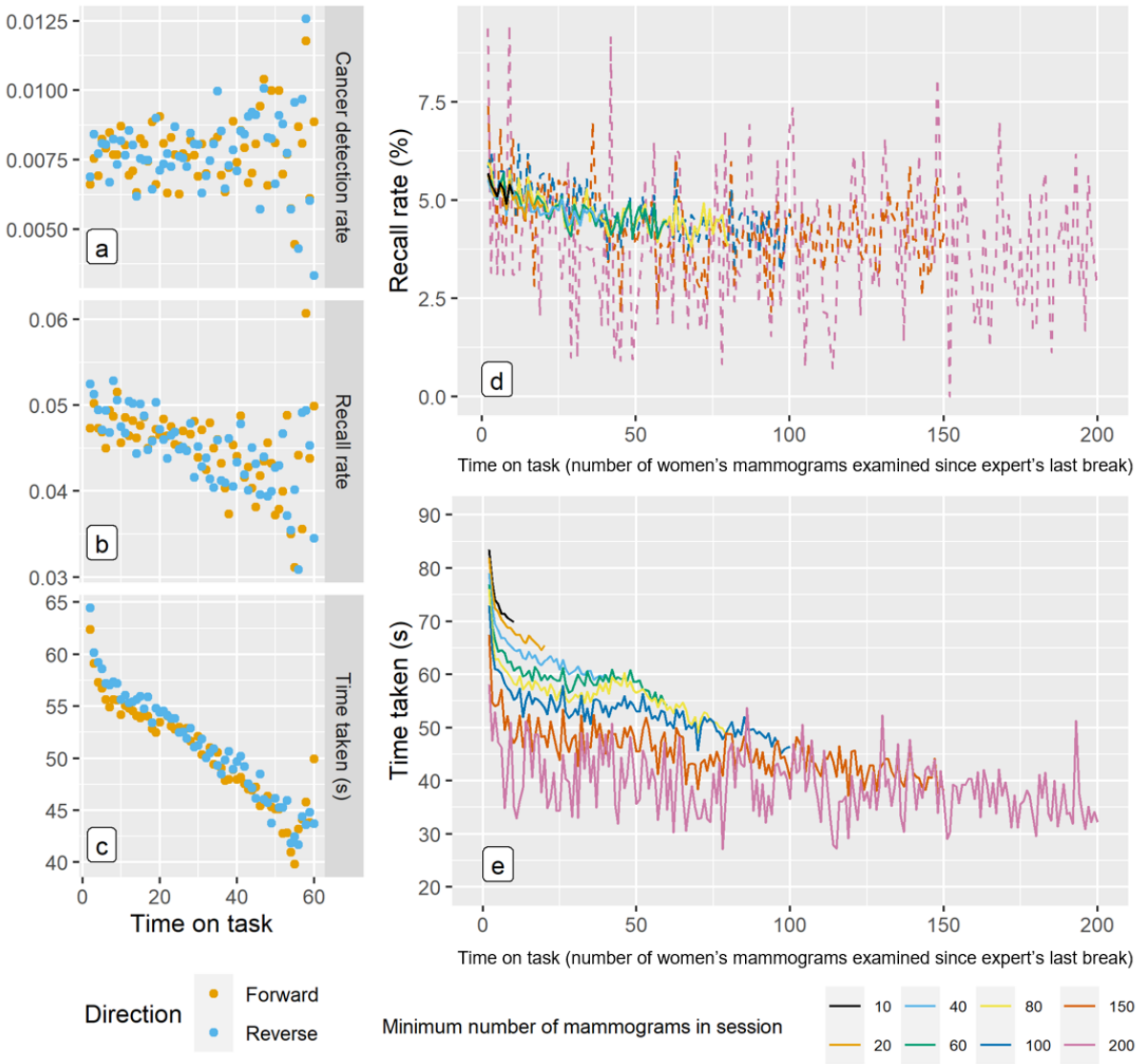

Figure 3. Robustness of results to individual woman level confounding (a/b/c) and expert level confounding (d/e). Cancer detection rate (a), recall rate (b) and speed of decision-making (mean time taken) (c) with time on task for the trial intervention arm, where women were organized into sessions and two experts examined them in the opposite order to one another. Data shown for 534,108 women's mammograms in the 'forward' direction and 524,908 in the 'reverse' direction (session position of mammograms read in intended order, plots truncated at position 60). Recall rate (d) and speed of decision-making (e) with time on task by minimum task length. Only reading sessions longer than the minimum task length (minimum number of mammograms in session) are included to remove potential confounding by different experts spending different time on task. After 20 minutes without a decision input the expert is considered to have taken a break and started a new session. Time on task represented by number of women's mammograms examined since the expert's last break.

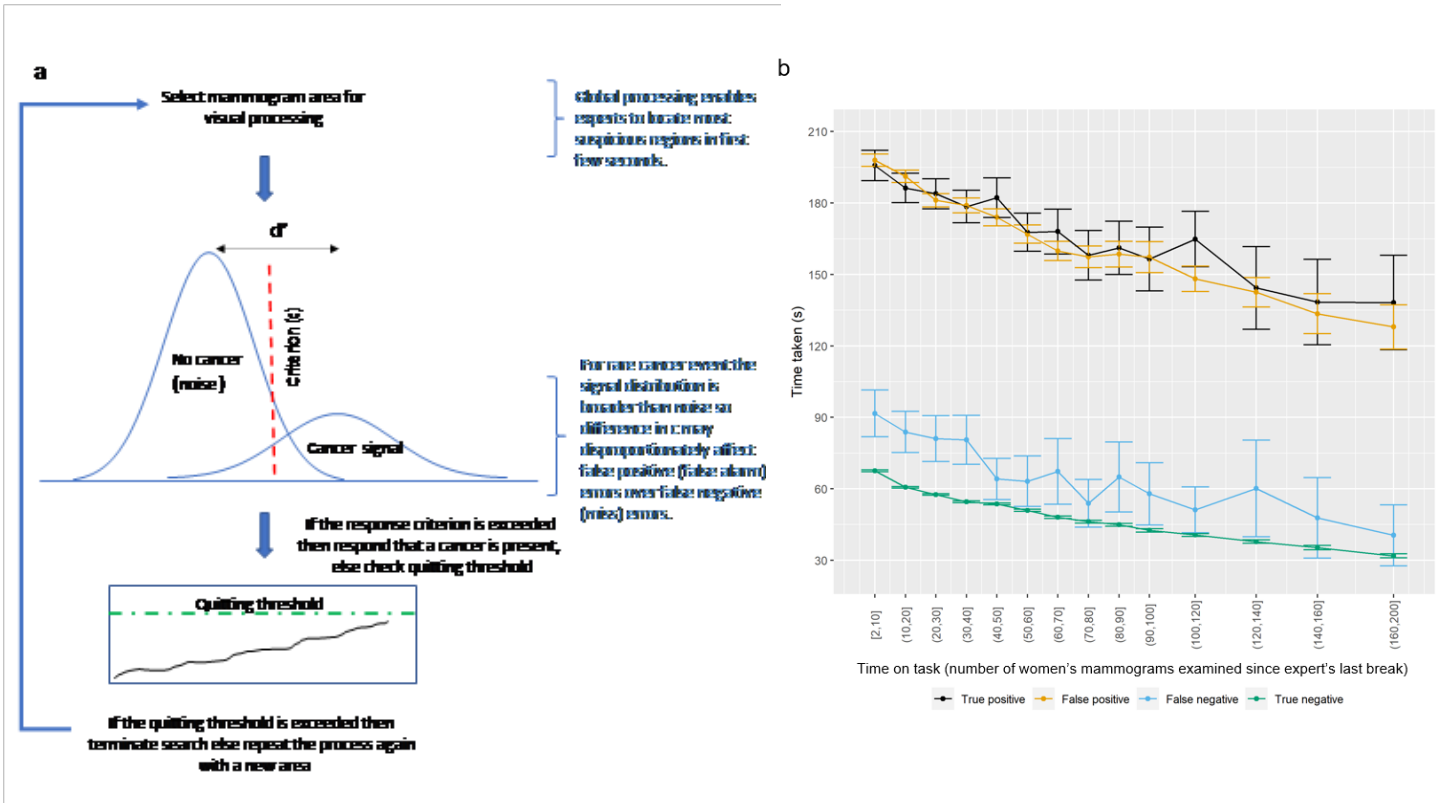

Figure 4 – (a) Multiple Decision Model proposed by Wolfe *et al* (2007)(39), with global processing(42). If the selected area contains a cancer and response falls to the right of the criterion decision line then the cancer would be detected, else it would be missed. This process would continue until the quitting threshold is reached, at which point search would terminate. With time on task, our data suggest the Quitting threshold (denoted by the green dotted line) would be lowered, and the response threshold (denoted by the red dotted line) would move to the right (so that experts are more conservative in their response to say there is a cancer). (b) our results for speed of reading (mean with 95%CI) with time on task (number of women's mammograms examined since the expert's last break), shown for each decision outcome (true positive, false positive, true negative or false negative outcome).

Supplementary Materials for:

Fatigue and Vigilance in Medical Experts Detecting Breast Cancer

Sian Taylor Phillips, David Jenkinson, Chris Stinton, Melina A Kunar, Derrick G Watson, Karoline Freeman, Alice Mansbridge, Matthew G Wallis, Olive Kearins, Sue Hudson, Aileen Clarke

Correspondence to: s.taylor-phillips@warwick.ac.uk

This PDF file includes:

Supplementary Text
Figs. S1 to S5
Tables S1 to S5

This supplement contains additional details for the study methods (section 1 figure S1), descriptive statistics (section 2 table S1), full model outputs for adjusted and unadjusted models (Section 3 tables S2-S5). Results for speed of reading are presented for median rather than mean, demonstrating no impact of excluding outliers on results (section 4 figure S2). Results for cancer detection rate, recall rate and speed of reading are presented both with and without cases which were examined out of the intended order, demonstrating this exclusion did not impact results (section 5, figure S3).

1. Supplementary Methods

Defining how many cases had been examined in each reading session

Experts examine mammograms in screening practice in predefined sessions. Each session consists of a full or half day for a screening mammography machine at one location. Sessions are created alphabetically by surname from lists of eligible women from general practitioners' databases. On the radiology software each session is opened separately and is displayed as a list of women for whom the expert sequentially decides whether to recall each for further tests. The computer software records any cases which were not examined in the sequential order assigned. After completing a session, experts often immediately start another session, which can be achieved simply and quickly with the click of a mouse. The primary analysis in this paper considers an expert to have kept working constantly even if switching session as long as they did not spend longer than the predefined times between making decisions for subsequent women.

To create the new sessions we put the mammograms examined by each expert into chronological order (using the date/time stamp). Individual experts read some mammograms as first expert and others as second expert. Thus, we separated the data from each mammogram into two records, one for the data from each expert. The dataset was sorted by centre, expert and date/time to give a list of cases read by each expert in chronological order

(regardless of whether it was as first expert or second expert). The difference between the time stamps of consecutive cases was calculated to give the time taken for that case by that expert.

The new sessions were then created (with new session numbers) by assigning each case to either the same session as the previous case if the time taken was less than the break definition, or to a new session if the time taken was greater than the break definition. This was done for each of the different values of the break definition (10,20, 60, 180 and 480 minutes).

It was possible that where mammograms from different original sessions have been combined into the same new session that the new session will contain some mammograms read as first expert and others as second expert. The dataset used for modelling contains all of the mammograms once, with their session position and outcomes taken from the first expert only. This also ensures that the mammograms were examined independently without consulting the other expert's decision.

The trial data included a field indicating whether the actual order that the mammogram was read in was the intended order. Mammograms that were not read in the intended order were excluded from the models, as they may be systematically different, for example occasionally difficult cases which were put aside for later review. Mammograms that were read in the first position of a new session (as defined by the different break definitions) were also excluded, as it is not possible to determine the time taken to read them.

It was also necessary to exclude data from centres where it was not possible to distinguish between different experts using the expert ID code in the dataset.

The distributions of the session position numbers used in the models are shown in Figure S1.

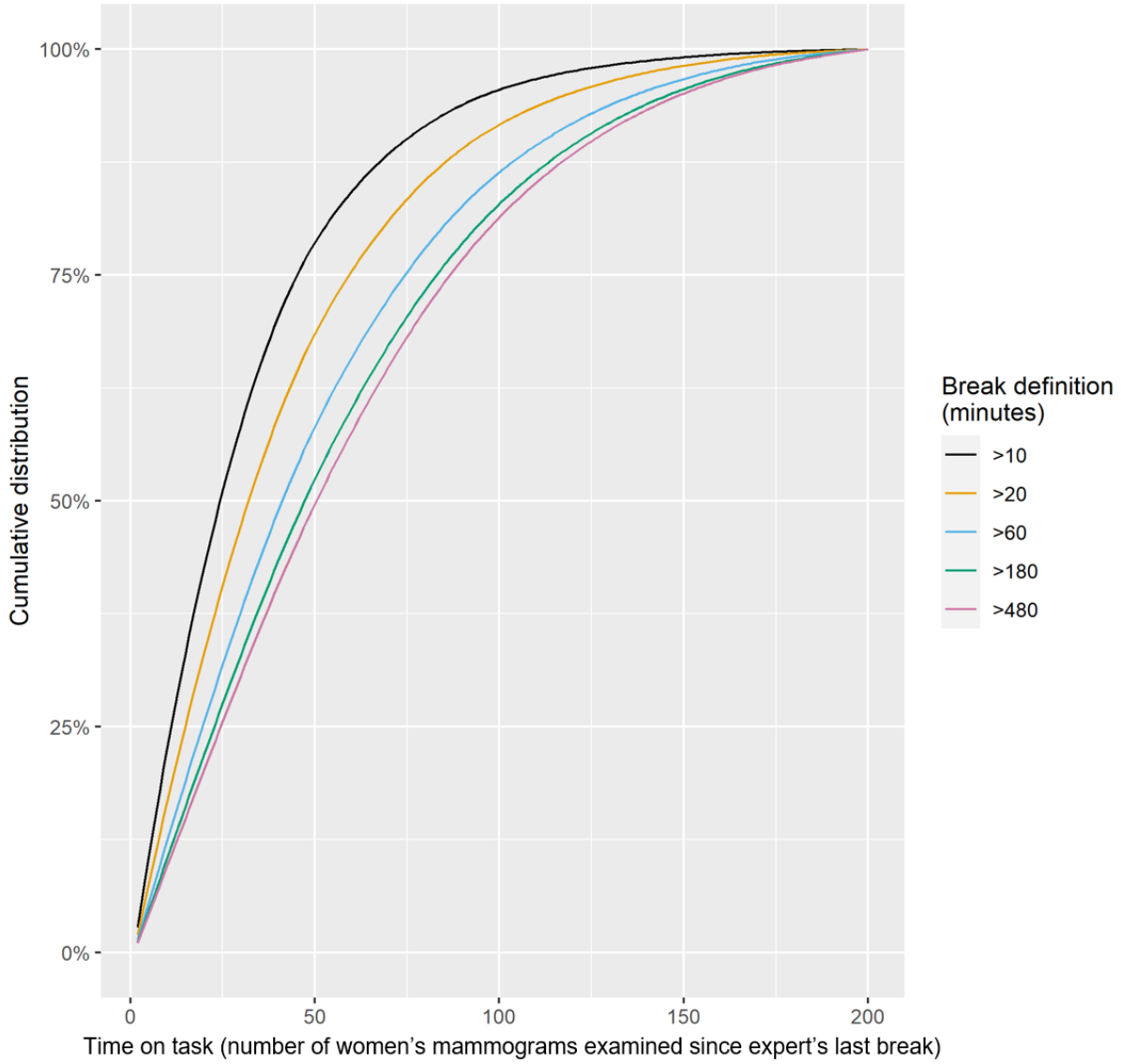

Figure S1. Cumulative distribution of women in the dataset by number of women's mammograms examined since experts last break (from 2 to 200). Includes only those examined in the intended order and with woman's age present in the dataset. Shown for all values of the break definition. The total number of mammograms read for each break definition is shown in row 2 of Table S1.

2. Supplementary tables - Descriptive statistics

The exclusion of mammograms from the dataset due to missing data to identify the expert (71,695) or due to the first expert not examining in the intended order (a further 52,886) reduced the dataset to 1,069,566 women's mammograms, of which 37 were a second set of mammograms from an individual woman.

The number of women included at each session position is dependent on the session definition and is given in figure S1. Using the first expert only, the recall rate was 4.8% and the cancer detection rate 7.4 per thousand women screened.

There were 410 experts in the study identified by their pseudonymised login at the computer system at each breast screening centre. Of these, only 360 pseudonymised codes were unique across the whole dataset, with the same pseudonymised code appearing at more than one centre. It was not possible to identify whether the same expert worked at two different breast screening centres or whether the same login was a coincidence, so we conservatively report only 360 readers.

Descriptive statistics of the three outcomes under the different thresholds are shown in Table S1.

Table S1. Descriptive statistics of the outcomes for each threshold.

	Definition of a break (time without inputting a decision into the software)				
	≥10 minutes	≥20 minutes	≥60minutes	≥180minutes	≥480 minutes
Number of women's mammograms (position 2-200)	1,037,145	1,042,597	1,040,970	1,035,758	1,034,031
Number of women's mammograms (position 2-200, age recorded)	1,037,140	1,042,592	1,040,965	1,035,753	1,034,026
Recall rate	4.63%	4.73%	4.77%	4.78%	4.79%
(women recalled/total women)	(47,993/1,307,140)	(49,344/1,042,592)	(49,610/1,040,965)	(49,548/1,035,753)	(49,508/1,034,026)
Cancer detection rate	0.726%	0.740%	0.745%	0.745%	0.746%
(women with cancer detected/total women)	(7,529/1,307,140)	(7,715/1,042,592)	(7,757/1,040,965)	(7,717/1,035,753)	(7,709/1,034,026)
Number of women's mammograms (position 2-200, age recorded, time between 1s and 600s inclusive)	1037135	1033433	1025888	1018158	1015296
Median time taken (inter-quartile range, s)	35 (19 - 69)	35 (19-69)	35 (20 – 69)	35 (20 – 69)	35 (20 – 70)
Interval cancers per 1,000 women	1.90	1.91	1.90	1.91	1.91
(women with interval cancer within 3 years / number of women)	(1,975/1,037,140)	(1,987/1,042,592)	(1,980/1,040,965)	(1,974/1,035,753)	(1,975/1,034,026)

3. Supplementary tables - Describing the models

The model coefficients for the main model adjusted for the woman's age and whether she had previously attended are given in table S2, with fitted values in table S3. The coefficients of an unadjusted but otherwise equivalent model are given in table S4, with fitted values in table S5.

Table S2. Model coefficients (with 95% confidence intervals) and random effect standard deviations for models for the three outcomes at the different break definitions. For recall and cancer detection the coefficients from their logistic models are shown as odds ratios (OR). For time taken to read the coefficients are shown on the linear scale. The models for recall and time taken used a linear basis spline for session position, with knots at positions 20 and 40. Instead of the model coefficients the gradient of those three lines is shown; as an OR for recall, and linearly for time taken. The gradient is over five session positions, rather than one. Age was standardised in the models. The coefficients shown in these tables have been adjusted (divided by the standard deviation, before conversion to the odds ratio scale) to show the effect of an increase of one year on the outcome. The intercept terms should be interpreted as the outcome at session position two for a mammogram of the mean age that is incident (not a woman's first mammogram). The model results for prevalent (woman's first mammogram) rather than incident are shown in the tables. The standard deviation of the random effects at each level are abbreviated to "RE SD". The cancer detection and recall results are shown to three decimal places (except for the cancer detection intercept and session position) and the time taken results to three significant digits.

	Definition of a break (time without inputting a decision into the software)				
	≥10 minutes	**≥20 minutes**	**≥60minutes**	**≥180minutes**	**≥480 minutes**
Cancer Detection Rate					
Intercept (Position 2)	0.00603	0.00631	0.00629	0.00628	0.00627
95% CI	(0.00572 - 0.00635)	(0.00599 - 0.00665)	(0.00597 - 0.00663)	(0.00595 - 0.00663)	(0.00594 - 0.00662)
Session Position (OR over five positions)	1.00070	0.99755	0.99899	0.99922	0.99937
95% CI	(0.99687 - 1.00454)	(0.99434 - 1.00078)	(0.99617 - 1.00182)	(0.99656 - 1.00189)	(0.99675 - 1.00199)
Age (OR)	1.052	1.052	1.052	1.052	1.052
95% CI	(1.048 - 1.056)	(1.048 - 1.056)	(1.048 - 1.056)	(1.048 - 1.056)	(1.048 - 1.056)
Prevalent (OR)	1.763	1.738	1.741	1.736	1.739
95% CI	(1.640 - 1.895)	(1.618 - 1.867)	(1.621 - 1.870)	(1.616 - 1.865)	(1.619 - 1.868)
RE SD - Centre	0.143	0.145	0.139	0.140	0.142
RE SD - Expert	0.100	0.096	0.097	0.098	0.097

Recall Rate					
Intercept (Position 2)	0.036	0.038	0.040	0.040	0.040
95% CI	(0.032 - 0.040)	(0.035 - 0.042)	(0.036 - 0.044)	(0.036 - 0.044)	(0.037 - 0.045)
Session Position 2-20 (OR over five positions)	0.974	0.973	0.970	0.973	0.970
95% CI	(0.964 - 0.984)	(0.962 - 0.984)	(0.958 - 0.982)	(0.960 - 0.986)	(0.957 - 0.983)
Session Position 20-40 (OR over five positions)	0.989	0.984	0.981	0.978	0.979
95% CI	(0.981 - 0.997)	(0.977 - 0.992)	(0.973 - 0.989)	(0.970 - 0.986)	(0.971 - 0.988)
Session Position 40-200 (OR over five positions)	0.994	0.993	0.994	0.995	0.994
95% CI	(0.991 - 0.997)	(0.991 - 0.995)	(0.992 - 0.995)	(0.993 - 0.996)	(0.993 - 0.996)
Age (OR)	1.004	1.005	1.005	1.005	1.005
95% CI	(1.003 - 1.006)	(1.003 - 1.006)	(1.003 - 1.006)	(1.003 - 1.006)	(1.003 - 1.006)
Prevalent (OR)	2.739	2.717	2.716	2.713	2.713
95% CI	(2.665 - 2.815)	(2.645 - 2.791)	(2.644 - 2.789)	(2.641 - 2.787)	(2.641 - 2.787)
RE SD - Centre	0.338	0.339	0.335	0.334	0.334
RE SD - Expert	0.303	0.298	0.299	0.298	0.297
Time taken to read (s)					
Intercept (Position 2)	72.3	73.3	74.9	76	76.5
95% CI	(72.1 - 72.6)	(73.1 - 73.5)	(74.6 - 75.2)	(75.6 - 76.5)	(76.2 - 76.8)
Session Position 2-20 (Over five positions)	-2.66	-2.58	-2.72	-2.91	-2.97
95% CI	(-2.71 - -2.61)	(-2.63 - -2.53)	(-2.77 - -2.67)	(-2.97 - -2.85)	(-3.02 - -2.91)
Session Position 20-40 (Over five positions)	0.0552	-0.101	-0.226	-0.234	-0.245
95% CI	(0.00471 - 0.106)	(-0.15 - -0.0522)	(-0.277 - -0.175)	(-0.291 - -0.176)	(-0.302 - -0.189)
Session Position 40-200 (Over five positions)	-0.243	-0.276	-0.295	-0.284	-0.284
95% CI	(-0.253 - -0.234)	(-0.283 - -0.269)	(-0.304 - -0.287)	(-0.293 - -0.275)	(-0.291 - -0.276)
Age (OR)	0.0473	0.047	0.0503	0.0498	0.048
95% CI	(0.036 - 0.0586)	(0.0357 - 0.0584)	(0.0387 - 0.0619)	(0.0386 - 0.0611)	(0.0373 - 0.0588)

Prevalent (OR)	1.57	1.59	1.62	1.51	1.46
95% CI	(1.4 - 1.74)	(1.38 - 1.79)	(1.42 - 1.81)	(1.34 - 1.68)	(1.31 - 1.61)
RE SD - Centre	28.9	29	28.8	28.8	28.7
RE SD - Expert	16.3	16.4	16.5	16.6	16.6
RE SD - Observation	1.25	1.25	1.25	1.24	1.24

Table S3. Fitted values for selected session positions from the models of different break definitions, with 95% confidence intervals, for all three outcomes. The session positions listed are chosen throughout the range and include the knot points used in the basis spline of session position used in the models, positions 20 and 40.

	Definition of a break (time without inputting a decision into the software)				
	≥10 minutes	**≥20 minutes**	**≥60minutes**	**≥180minutes**	**≥480 minutes**
Cancer Detection Rate (%) at session position					
2	0.679	0.708	0.707	0.705	0.704
95% CI	(0.646 - 0.712)	(0.675 - 0.744)	(0.673 - 0.743)	(0.670 - 0.742)	(0.669 - 0.741)
20	0.680	0.702	0.704	0.703	0.703
95% CI	(0.651 - 0.711)	(0.672 - 0.734)	(0.673 - 0.737)	(0.672 - 0.736)	(0.671 - 0.736)
40	0.682	0.696	0.701	0.701	0.701
95% CI	(0.654 - 0.712)	(0.667 - 0.725)	(0.673 - 0.732)	(0.672 - 0.732)	(0.672 - 0.732)
60	0.684	0.689	0.699	0.699	0.699
95% CI	(0.652 - 0.717)	(0.660 - 0.719)	(0.670 - 0.729)	(0.670 - 0.729)	(0.671 - 0.729)
100	0.688	0.675	0.693	0.695	0.696
95% CI	(0.644 - 0.735)	(0.638 - 0.715)	(0.659 - 0.729)	(0.662 - 0.729)	(0.664 - 0.729)
150	0.693	0.659	0.686	0.689	0.692
95% CI	(0.628 - 0.765)	(0.608 - 0.715)	(0.640 - 0.736)	(0.646 - 0.736)	(0.649 - 0.737)
200	0.698	0.643	0.679	0.684	0.687
95% CI	(0.610 - 0.798)	(0.576 - 0.718)	(0.618 - 0.746)	(0.627 - 0.746)	(0.631 - 0.748)
Recall rate (%) at session position					
2	4.397	4.656	4.839	4.891	4.947
95% CI	(3.990 - 4.843)	(4.230 - 5.123)	(4.391 - 5.330)	(4.436 - 5.390)	(4.486 - 5.454)
20	4.017	4.236	4.363	4.451	4.459
95% CI	(3.646 - 4.424)	(3.851 - 4.659)	(3.963 - 4.801)	(4.043 - 4.898)	(4.050 - 4.907)
40	3.844	3.990	4.057	4.087	4.116

95% CI	(3.489 - 4.234)	(3.628 - 4.386)	(3.689 - 4.460)	(3.717 - 4.492)	(3.744 - 4.523)
60	3.759	3.887	3.958	4.002	4.028
95% CI	(3.414 - 4.138)	(3.537 - 4.271)	(3.601 - 4.349)	(3.641 - 4.396)	(3.666 - 4.424)
100	3.595	3.690	3.767	3.836	3.858
95% CI	(3.257 - 3.966)	(3.354 - 4.059)	(3.426 - 4.142)	(3.490 - 4.215)	(3.511 - 4.239)
150	3.399	3.458	3.541	3.638	3.656
95% CI	(3.052 - 3.784)	(3.125 - 3.824)	(3.208 - 3.907)	(3.300 - 4.009)	(3.318 - 4.027)
200	3.213	3.239	3.328	3.450	3.463
95% CI	(2.843 - 3.631)	(2.900 - 3.617)	(2.996 - 3.696)	(3.113 - 3.822)	(3.128 - 3.834)
Time taken to read (s) at session position					
2	72.7	73.7	75.3	76.4	76.8
95% CI	(72.4 - 72.9)	(73.4 - 73.9)	(74.9 - 75.6)	(75.9 - 76.8)	(76.5 - 77.1)
20	63.1	64.4	65.5	65.9	66.1
95% CI	(62.8 - 63.4)	(64.1 - 64.6)	(65.1 - 65.9)	(65.4 - 66.4)	(65.8 - 66.4)
40	63.3	64.0	64.6	64.9	65.1
95% CI	(63.0 - 63.6)	(63.7 - 64.2)	(64.2 - 64.9)	(64.4 - 65.4)	(64.8 - 65.4)
60	62.3	62.9	63.4	63.8	64.0
95% CI	(62.1 - 62.6)	(62.6 - 63.1)	(63.0 - 63.7)	(63.3 - 64.3)	(63.7 - 64.3)
100	60.4	60.6	61.0	61.5	61.7
95% CI	(60.1 - 60.7)	(60.4 - 60.9)	(60.7 - 61.4)	(61.0 - 62.0)	(61.4 - 62.0)
150	58.0	57.9	58.1	58.7	58.9
95% CI	(57.7 - 58.3)	(57.6 - 58.1)	(57.7 - 58.4)	(58.2 - 59.2)	(58.6 - 59.2)
200	55.5	55.1	55.1	55.8	56.0
95% CI	(55.2 - 55.9)	(54.9 - 55.4)	(54.7 - 55.5)	(55.3 - 56.4)	(55.7 - 56.4)

Table S4. Model coefficients (with 95% confidence intervals) and random effect standard deviations for models for the three outcomes at the different break definitions for sessions, which have not been adjusted for age and prevalence status. For recall and cancer detection the coefficients from their logistic models are shown as odds ratios (OR). For time taken to read the coefficients are shown on the linear scale. The models for recall and time taken used a linear basis spline for session position, with knots at positions 20 and 40. Instead of the model coefficients the gradient of those three lines is shown; as an OR for recall, and linearly for time taken. The gradient is over five session positions, rather than one. The intercept terms should be interpreted as the outcome at session position two. The standard deviation of the random effects at each level are abbreviated to "RE SD". The cancer detection and recall results are shown to three decimal places (except for the cancer detection intercept and session position) and the time taken results to three significant digits.

	Definition of a break (time without inputting a decision into the software)				
	≥10 minutes	**≥20 minutes**	**≥60minutes**	**≥180minutes**	**≥480 minutes**
Cancer Detection Rate					
Intercept (Position 2)	0.00712	0.00743	0.00742	0.00740	0.00739
95% CI	(0.00679 - 0.00747)	(0.00709 - 0.00779)	(0.00707 - 0.00779)	(0.00704 - 0.00777)	(0.00703 - 0.00777)
Session Position (OR over five positions)	1.00044	0.99752	0.99893	0.99919	0.99936
95% CI	(0.99662 - 1.00427)	(0.99431 - 1.00074)	(0.99611 - 1.00175)	(0.99653 - 1.00186)	(0.99675 - 1.00198)
RE SD – Centre	0.143	0.145	0.140	0.141	0.143
RE SD – Expert	0.096	0.092	0.093	0.094	0.093
Recall Rate					
Intercept (Position 2)	0.048	0.051	0.053	0.053	0.054
95% CI	(0.044 - 0.053)	(0.046 - 0.056)	(0.048 - 0.058)	(0.048 - 0.059)	(0.049 - 0.059)
Session Position 2-20 (OR over five positions)	0.968	0.966	0.963	0.966	0.963
95% CI	(0.958 - 0.978)	(0.956 - 0.977)	(0.951 - 0.975)	(0.953 - 0.979)	(0.951 - 0.977)
Session Position 20-40 (OR over five positions)	0.989	0.985	0.982	0.98	0.981

95% CI	(0.981 - 0.997)	(0.978 - 0.993)	(0.974 - 0.990)	(0.972 - 0.988)	(0.973 - 0.990)
Session Position 40-200 (OR over five positions)	0.996	0.994	0.995	0.995	0.995
95% CI	(0.993 - 0.999)	(0.992 - 0.997)	(0.993 - 0.996)	(0.993 - 0.997)	(0.993 - 0.996)
RE SD - Centre	0.336	0.337	0.334	0.333	0.333
RE SD - Expert	0.308	0.302	0.303	0.301	0.300
Time taken to read (s)					
Intercept	72.7	73.7	75.3	76.4	76.8
95% CI	(72.5 - 72.9)	(73.5 - 73.9)	(75 - 75.6)	(76.1 - 76.6)	(76.4 - 77.2)
Session Position 2-20 (Over five positions)	-2.66	-2.59	-2.72	-2.92	-2.97
95% CI	(-2.7 - -2.62)	(-2.63 - -2.54)	(-2.78 - -2.67)	(-2.97 - -2.86)	(-3.02 - -2.92)
Session Position 20-40 (Over five positions)	0.053	-0.101	-0.228	-0.234	-0.246
95% CI	(0.0102 - 0.0957)	(-0.146 - -0.056)	(-0.283 - -0.172)	(-0.289 - -0.18)	(-0.304 - -0.188)
Session Position 40-200 (Over five positions)	-0.243	-0.276	-0.295	-0.284	-0.284
95% CI	(-0.25 - -0.237)	(-0.284 - -0.269)	(-0.303 - -0.287)	(-0.292 - -0.276)	(-0.292 - -0.275)
RE SD - Centre	28.9	29	28.8	28.8	28.7
RE SD - Expert	16.3	16.4	16.5	16.6	16.6
RE SD - Observation	1.25	1.25	1.25	1.24	1.24

Table S5. Fitted values for selected session positions from the models of different thresholds not adjusted for age and prevalence status, with 95% confidence intervals, for all three outcomes. The session positions listed are chosen throughout the range and include the knot points used in the basis spline of session position used in the recall and time taken models, positions 20 and 40.

	Definition of a break (time without inputting a decision into the software)				
	≥10 minutes	**≥20 minutes**	**≥60minutes**	**≥180minutes**	**≥480 minutes**
Cancer Detection Rate (%) at session position					
2	0.712	0.743	0.742	0.740	0.739
95% CI	(0.679 - 0.747)	(0.709 - 0.779)	(0.707 - 0.779)	(0.704 - 0.777)	(0.703 - 0.777)
20	0.713	0.737	0.739	0.738	0.737
95% CI	(0.684 - 0.744)	(0.706 - 0.768)	(0.708 - 0.772)	(0.705 - 0.771)	(0.705 - 0.771)
40	0.715	0.729	0.736	0.735	0.735
95% CI	(0.686 - 0.745)	(0.701 - 0.759)	(0.707 - 0.766)	(0.705 - 0.766)	(0.706 - 0.767)
60	0.716	0.722	0.733	0.733	0.734
95% CI	(0.684 - 0.750)	(0.693 - 0.753)	(0.704 - 0.763)	(0.704 - 0.763)	(0.705 - 0.764)
100	0.718	0.708	0.727	0.728	0.730
95% CI	(0.673 - 0.767)	(0.670 - 0.748)	(0.692 - 0.763)	(0.695 - 0.763)	(0.697 - 0.764)
150	0.722	0.691	0.719	0.722	0.725
95% CI	(0.654 - 0.796)	(0.637 - 0.749)	(0.671 - 0.770)	(0.677 - 0.770)	(0.681 - 0.772)
200	0.725	0.674	0.711	0.717	0.721
95% CI	(0.634 - 0.828)	(0.604 - 0.752)	(0.648 - 0.781)	(0.657 - 0.781)	(0.663 - 0.784)
Recall rate (%) at session position					
2	4.806	5.082	5.287	5.338	5.394
95% CI	(4.362 - 5.293)	(4.618 - 5.590)	(4.798 - 5.822)	(4.844 - 5.879)	(4.896 - 5.940)
20	4.296	4.518	4.648	4.740	4.749
95% CI	(3.899 - 4.732)	(4.106 - 4.969)	(4.222 - 5.115)	(4.306 - 5.215)	(4.315 - 5.224)
40	4.115	4.270	4.343	4.387	4.420

95% CI	(3.734 - 4.533)	(3.882 - 4.694)	(3.948 - 4.775)	(3.991 - 4.820)	(4.023 - 4.855)
60	4.049	4.180	4.253	4.303	4.333
95% CI	(3.676 - 4.457)	(3.803 - 4.593)	(3.868 - 4.673)	(3.917 - 4.727)	(3.945 - 4.757)
100	3.919	4.007	4.078	4.141	4.163
95% CI	(3.552 - 4.324)	(3.641 - 4.408)	(3.707 - 4.483)	(3.768 - 4.550)	(3.789 - 4.571)
150	3.764	3.800	3.869	3.947	3.959
95% CI	(3.381 - 4.188)	(3.435 - 4.202)	(3.505 - 4.269)	(3.581 - 4.348)	(3.595 - 4.359)
200	3.614	3.603	3.670	3.761	3.766
95% CI	(3.200 - 4.078)	(3.227 - 4.021)	(3.304 - 4.075)	(3.394 - 4.165)	(3.402 - 4.166)
Time taken to read (s) at session position					
2	72.7	73.7	75.3	76.4	76.8
95% CI	(72.5 - 72.9)	(73.5 - 73.9)	(75.0 - 75.6)	(76.1 - 76.6)	(76.4 - 77.2)
20	63.1	64.4	65.5	65.9	66.1
95% CI	(62.9 - 63.3)	(64.1 - 64.6)	(65.1 - 65.8)	(65.5 - 66.2)	(65.7 - 66.6)
40	63.3	64.0	64.6	64.9	65.1
95% CI	(63.1 - 63.5)	(63.7 - 64.2)	(64.2 - 64.9)	(64.6 - 65.2)	(64.7 - 65.5)
60	62.3	62.9	63.4	63.8	64.0
95% CI	(62.1 - 62.6)	(62.6 - 63.1)	(63.1 - 63.7)	(63.5 - 64.1)	(63.6 - 64.4)
100	60.4	60.6	61.0	61.5	61.7
95% CI	(60.2 - 60.6)	(60.4 - 60.9)	(60.7 - 61.4)	(61.3 - 61.8)	(61.3 - 62.2)
150	58.0	57.9	58.1	58.7	58.9
95% CI	(57.8 - 58.2)	(57.6 - 58.1)	(57.7 - 58.4)	(58.4 - 59.0)	(58.4 - 59.4)
200	55.5	55.1	55.1	55.8	56.0
95% CI	(55.3 - 55.8)	(54.8 - 55.4)	(54.7 - 55.5)	(55.6 - 56.1)	(55.5 - 56.6)

66

4. Median speed of reading

The primary measure of time taken to examine each case was a mean, with cases taking longer than 10 minutes excluded so that the mean was not overly influenced by the tail of the distribution. The median time taken with no exclusions is shown in figure S2. This demonstrates the same pattern of decreasing time taken per case with increasing time on task.

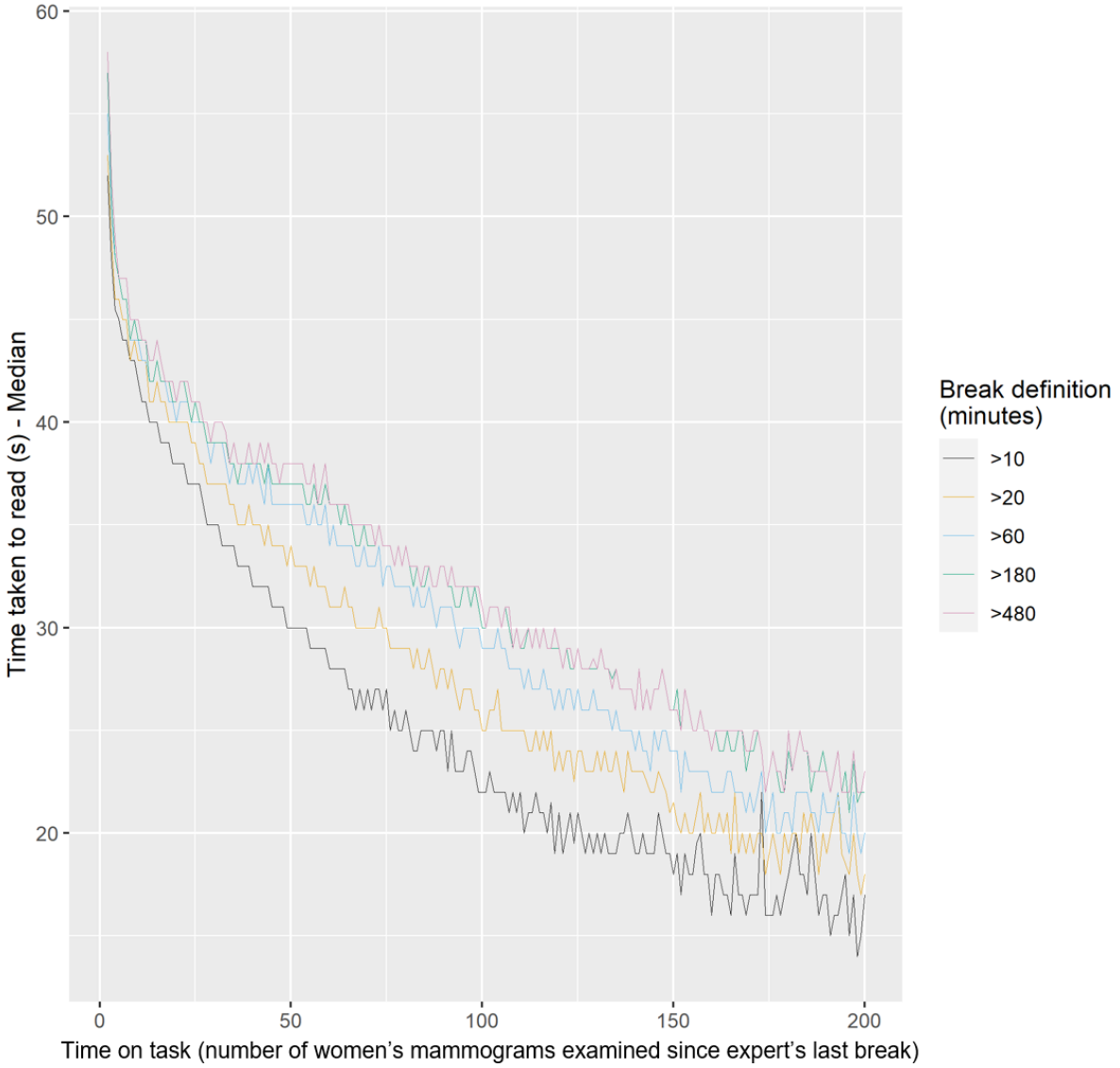

Figure S2. Median time taken to read for each position in a reading session, by break definition. Mammograms read in order only, excluding those with time taken to read of zero, but with no upper limit on time taken.

5. Models including cases examined out of intended order

The main models exclude cases which were not examined in the intended order, so that experts coming back to more difficult cases later could not bias results. Models were repeated including these cases examined out of the intended order and it did not affect results, as shown in figure S3.

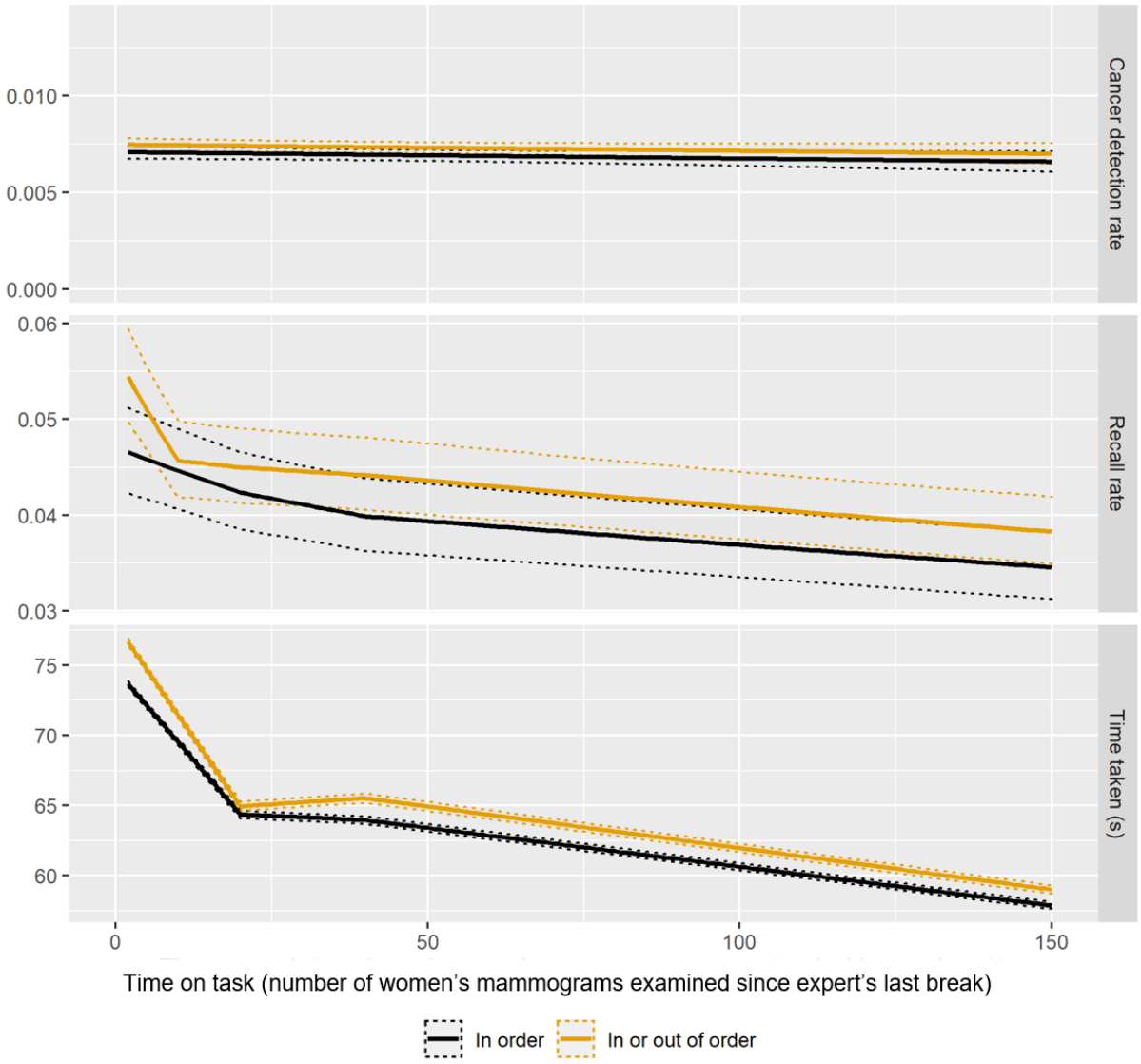

Figure S3. Model fitted values (shown as a solid line with 95% confidence intervals as dotted lines) including (orange) and excluding (black) women's mammograms examined in a different order to that intended. Performance metrics with time on task: cancer detection rate calculated as the proportion of women in which cancer was detected by the expert; recall rate calculated as the proportion of women that the expert indicated required recall for further tests; and mean time

taken to examine each woman's mammograms. Time on task is represented by the number of women's mammograms examined consecutively without a break. Break defined as 20 minutes without inputting a decision on the computer. Models were adjusted for women's age and whether she has previously attended screening, with clustering for expert and screening centre.

6. Break duration

To investigate the possible interaction between the length of the reader's breaks and the vigilance decrement/ improvement, outcomes for sessions starting after a short break, a moderate break, or a long break were plotted. This demonstrates that specificity and criterion are lower after a long break of >12hours, compared to after a short break of <1hour.

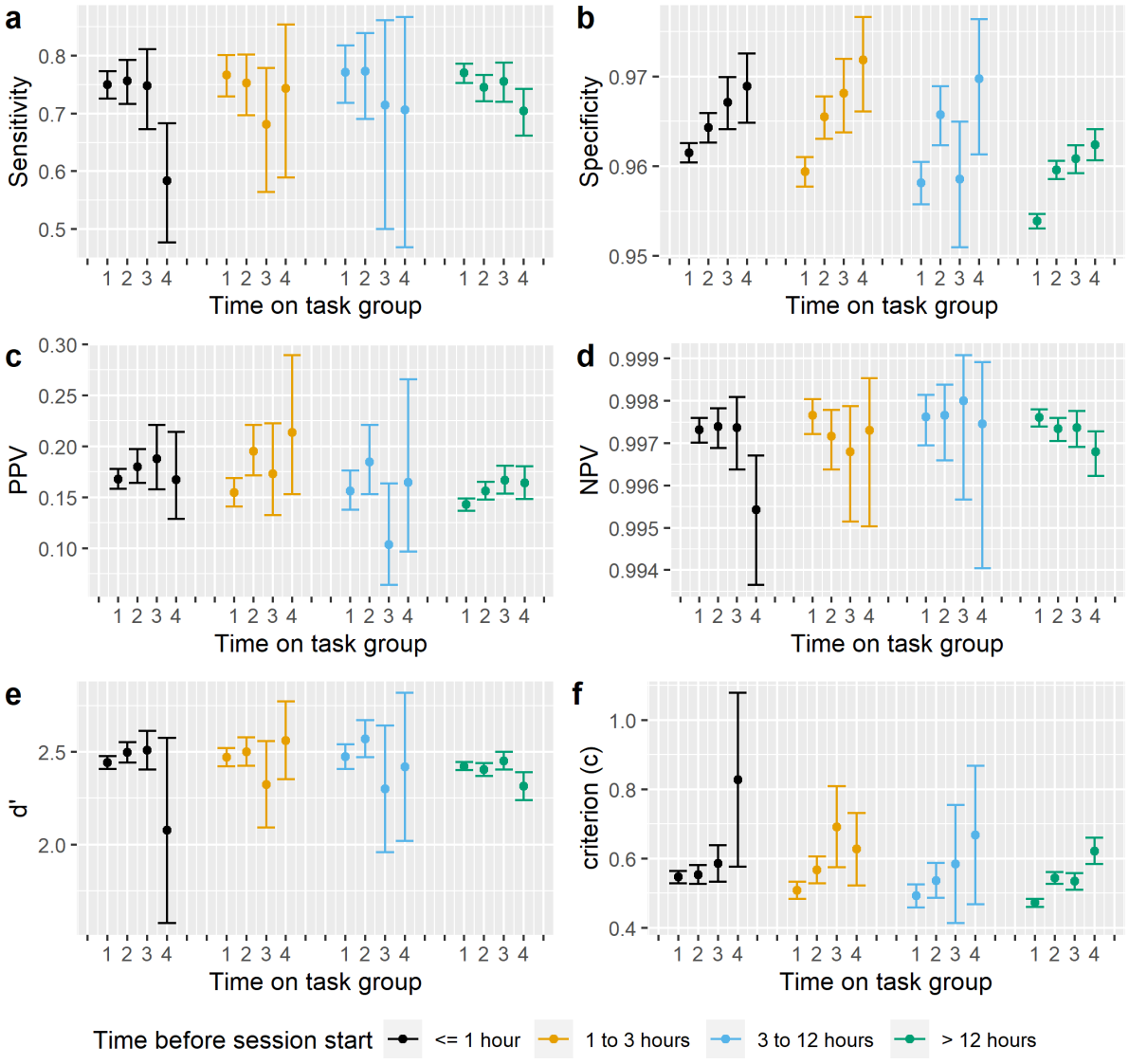

Figure S4. Outcomes for sessions starting after a short break of <1hour, a moderate break of 1 to 3 or 3 to 12 hours, or a long break of >12 hours which represents the next working day. The session is considred ended after a break without entering an opinion of more than 20 minutes in all definitions.

7. Time taken to read

To investigate the difference between the mean and median time taken to examine each woman's mammogram, the time taken to read was plotted using a 20-minute threshold:

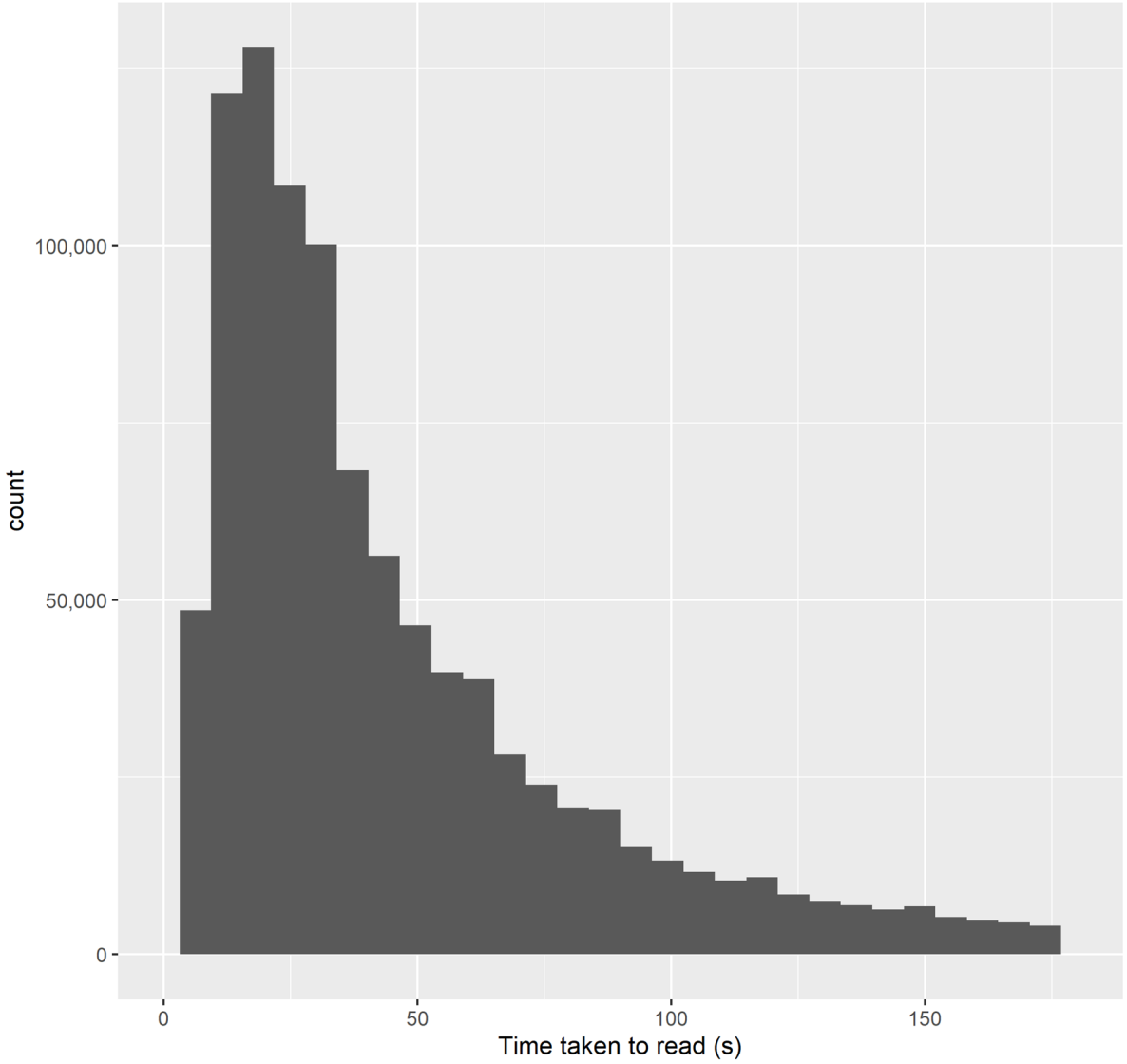

Figure S5. Histogram of time taken to read, using 20 minute threshold, horizontal axis limited to 180s.