Optimum screening mammography reading volumes: evidence from the NHS Breast Screening Programme

Abstract

Objectives: Minimum case load standards for professionals examining breast screening mammograms vary from 480 (US) to 5000 (Europe). We measured the relationship between number of women's mammograms examined per year and reader performance.

METHODS: We extracted routine records from the English NHS Breast Screening Programme for readers examining between 1000 and 45,000 mammograms between April 2014 and March 2017.We measured the relationship between volume of cases read and screening performance (cancer detection rate, recall rate, positive predictive value of recall (PPV) and discrepant cancers) using linear logistic regression. We also examined the effect of reader occupational group on performance. .

RESULTS:

759 eligible mammography readers (445 consultant radiologists, 235 radiography advanced practitioners, 79 consultant radiographers) examined 6.1 million women's mammograms during the study period. PPV increased from 12.9% to 14.4% to 17.0% for readers examining 2000, 5000 and 10000 cases per year respectively. This was driven by decreases in recall rates from 5.8 to 5,3 to 4.5 with increasing volume read, and no change in cancer detection rate (from 7.6 to 7.6 to 7.7). There was no difference in cancer detection rate with reader occupational group. Consultant radiographers had higher recall rate and lower PPV compared to radiologists (OR 1.105, p= 0.012; OR 0.874, p=0.002, unadjusted)

CONCLUSION: Positive predictive value of screening increases with total volume of cases examined per reader, through decreases in numbers of cases recalled with no concurrent change in numbers of cancers detected.

Keywords

Breast, breast cancer, mass screening, mammography

Key Points

- In the English Breast Screening Programme readers who examined a larger number of cases per year had higher positive predictive value, because they recalled fewer women for further tests but detected the same number of cancers
- Reader type did not affect cancer detection rate, but consultant radiographers had higher recall rate and lower positive predictive value than consultant radiologists, although this was not adjusted for length of experience

Abbreviations

BSIS-Breast Screening Information System

CDR-cancer detection rate

IQR-inter quartile range

KC62-National Health Service Breast Screening Programme Central Return Data Set

NHSBSP-National Health Service Breast Screening programme

OR-odds ratio

PPV-Positive predictive value

Introduction

Population based breast screening programmes aim to maximise the detection of breast cancer whilst minimising harm caused by false positive recalls.

Standards for annual mammography reading volumes vary internationally ranging from 480 in the USA to 2000 in Australia and 5000 in Europe [1, 2, 3]. There is no international consensus on the optimum reading volume. Since the inception of the National Health Service Breast Screening Programme (NHSBSP) it has been a requirement that all film readers report 5000 cases annually. This figure was originally based on expert opinion rather than evidence that it would affect key performance indices [4].

The problem with available research is that the majority has not been based on "real life" but rather assessing performance in test sets and correlating it with a reader reported volume. Experimental conditions can affect human behaviour and there has been criticism that test-set based performance research may suffer from a "laboratory effect" and not be a true reflection of real-life performance. In addition to the fact that test sets are usually enriched with cancer cases reading decisions may be altered because the reader knows that decisions they make in the test environment will have no impact on patient care [5].

Most of the real life studies that have been performed have been in countries with low reader volumes. These studies have usually found that higher-volume readers have lower false-positive rates with no sensitivity difference [6, 7, 8].

Two previous UK studies have examined the relationship between performance measures and volume of mammograms read looking at regional data in the East Midlands [9] and Scotland [10]. The results differed with Cornford et al [9] demonstrating lower recall rates in high volume readers and Duncan et al [10] finding no differences in any of the observed performance measures for different reader volumes. However both of these studies had relatively low numbers of readers with a total of 37 readers in each.

Recently Hoff et al published national data on the influence of mammography reader volume on radiologists' performance from Breast Screen Norway [11] They suggested annual reading volumes between 4000 and 10,000 were optimal with the main effect on specificity with fewer false positives at higher volumes. Cancer detection rates were stable up to 10,000 reads, with a trend to decrease thereafter.

The present study has examined national data from the NHSBSP in England over a 3 year period where all mammography has been digital. The primary aim was to measure the relationship between number of mammograms examined per year and reader performance. A secondary aim was to examine performance between different occupational groups. This is important both in terms of ensuring quality of the programme but also workforce planning.

Methods

The Screening Programme

The NHSBSP routinely invites women for 2 view digital mammography between 50 and up to their 71st birthday. There is currently a randomised age extension trial such that some women between 47 and 49 and 71 and 73 are also invited. Breast screening is offered on a 3 yearly basis. Women aged 71 and over can self-refer for breast screening every three years. All screening mammograms were included in this study, but mammograms examined as part of the symptomatic service, or screening very high risk women were excluded. Mammograms are routinely double read with either consensus or arbitration for discordant cases i.e if both readers recall a case then the woman is recalled, if only one reader recalls a case then the decision to recall is made by either some type of group discussion (consensus) or independently by a third reader (arbitration). Some centres also arbitrate cases recalled by both readers. NHSBSP guidelines state that mammography readers are required to read 5000 cases annually [4]. Mammography readers within the NHS come from different occupational groups either radiographers (advanced practitioners or consultant radiographers) or doctors (consultant radiologists and breast clinicians). Advanced practitioners have been reporting screening mammograms within the NHSBSP for approximately 20 years since the introduction of the NHS Cancer Plan in 2000 [12]. This launched new ways of working to streamline cancer services around the needs of the patient; through extending the roles of radiographers, nurses and other staff. Advanced practitioners are radiographers who have successfully completed a university accredited training course in mammography interpretation at a national breast imaging training centre. A consultant radiographer in addition to this training will have also gained the skills necessary to lead breast cancer assessment including ultrasound and image guided biopsy. In addition to consultant radiologists specialising in breast imaging there are also non radiologists specifically trained in all aspects of breast imaging and assessment termed "breast clinicians". For the purposes of this paper consultant radiologists and breast clinicians have been grouped together. Protocols for mammography screen reading are standardised across occupational groups. The method of resolving discordant decisions between readers is decided at unit rather than national level. It is dependent on local factors including staff numbers and staff experience. The most common methods are arbitration by a third reader or group consensus (with varying group sizes). Arbitration can be undertaken by staff from any occupational group but the requirements are greater than for the first and second read, and more commonly includes experienced radiologists. Guidance on who can undertake arbitration was published by Public Health England in 2016 [13].

A summary of the training undertaken by the different occupational groups and the experience necessary to arbitrate is summarised in supplementary material.

The Database and Inclusion Criteria

Recently, the English Breast Screening Information System (BSIS) which provides national and service level performance monitoring statistics for the NHSBSP has produced individual performance data in clinical practice over rolling three-year periods. BSIS collates the data from each screening service and aggregates at a national level for those readers with recorded activity at more than one screening service. The identification of such individuals is possible as BSIS requires the unique professional registration identifier for each reader with uploaded data. BSIS receives feeds using a standardised report from all NHS breast screening services therefore achieving a comprehensive dataset of practice.

All screening mammograms were included in this study, but mammograms examined as part of the symptomatic service, or screening high risk women with a family history of breast cancer were excluded. We only included decisions made as the first reader. In many centres, the second reader is not blinded to the opinion of the first reader and so may be biased.

Report outputs from each service were collected for the three-year period 01 April 2014 to 31 March 2017. Extracting the data in a three-year period reduced the probability of including a woman twice, as the programme is designed to invite women for screening once every three years. These outputs were loaded to a database specifically designed to hold and query these data. The query design allowed for further aggregation of those readers with recorded activity at more than one service. Data were pseudonymised so the analyst could not identify any individual. We excluded readers who had examined fewer than 1000 or more than 45,000 cases in total (as either first or second reader) between 1st April 2014 and 31st March 2017, because these extreme cases represented outliers in the analysis with the potential for high influence on the model..

Definition of variables

The unit of analysis was the individual reader. The volume of cases read was defined as the total number of screening cases read by each individual within the date limits as either first or second reader. For all outcomes only cases examined as the first reader were included. Invasive and non-invasive breast malignancies are included. The outcomes were defined as follows:

The recall rate was defined as the proportion of cases examined as first reader which the individual recalled to assessment.

The cancer detection rate (CDR) was defined as the number of histologically proven cancers detected per thousand women examined as the first reader.

The positive predictive value (PPV) was defined as the proportion of cases read that the individual recalled as first reader, which had histologically proven cancer.

The discrepant rate was defined as the number of cases which were not recalled by the first reader, but were found to have histologically proven cancer through recall by the other reader, expressed per 1000 women examined as first reader.

Finally, occupational group was also extracted and categorised as Advanced Practitioner, Consultant Radiographer and Radiologist. There was no follow up to subsequent interval or symptomatic cancers detection, as these cancers are not linked to individual readers within the database.

<u>Statistical Analysis</u>

Linear logistic regression models were used to examine the relationship between readers' 3 year total volume and the outcomes (PPV, recall rate, CDR and discrepant rate). Other modelling approaches were investigated, including fitting cubic splines, but the linear model was preferred in the analyses of all four outcomes due to superior model fit (lower Akaike information criterion and Bayesian information criterion). A dispersion parameter, which was estimated using the deviance of the model, was applied to adjust for over-dispersion. There was no adjustment for covariates. Information on the years of breast screening experience of each reader was not available, so we were unable to adjust for this. The effect of reader role (i.e. occupational group) on the measures is examined by adding reader role as an additional predictor to the earlier linear logistic regression models for each measure. Interaction terms between reader role and volume were examined for each model. We did not analyse discrepant rate by occupational group, as this would be heavily influenced by arbitration decision, which uses group consensus at some centres, and third reader (preferentially from the radiologist professional group) at others, and so such analysis may be biased in favour of the more senior consultant radiologists who are more influential at arbitration.

This analysis of de-identified data is classified as audit as part of the quality assurance function of the breast screening programme by Public Health England and therefore did not require ethical committee review.

Results

In the three years between 1st April 2014 and 31st March 2017, 863 readers examined a total of 6.4 million women's mammograms in the English NHS Breast Screening Programme. Each woman's mammograms were examined twice by two readers. We excluded 104 readers from the analysis: 93 because they read less than 1000 mammograms during the study period and 11 because they read more than 45,000 mammograms during the study period. Of the 759 included readers, 322 readers (42%) read less than the recommended 5000 mammograms per year as the first or second reader, and 158 readers (21%) read less than 1500 per year as first reader. The final dataset analysed included 759 readers who examined 6.1 million women's mammograms.

There were 445 consultant radiologists who examined a median of 16,519 cases each (IQR 10,128 to 21,024), 235 radiography advanced practitioners who examined a median of 16,077 cases each (IQR 8,971 to 20,811) and 79 consultant radiographers who examined a median of 16,507 cases each (IQR 12,380 to 20,000). The BSIS dataset does not include the demographics of women screened; however the National NHSBSP Central Return Data Set (KC62) for the same time period, for the same 6,466,565 women show a median age of 58.3, with 18.8% of these at their first breast screening appointment.

Relationship between volume of cases read and screening performance measures

Readers who examined a greater number of cases per year had a higher PPV (OR 1.013 for 1000 extra cases examined within 3 years, 95%CI 1.010 to 1.017, $p<0.001$), see Figure 1. The PPV increased from 12.9% for a reader examining 2000 cases per year to 19.9% for a reader examining 15,000 cases per year, see Table 1. This increase in PPV was driven by decreases in recall rate with increasing volume of cases read (OR 0.989, 95%CI 0.986 to 0.992, $p<0.001$). Recall rate decreased from 5.8% for a reader examining 2000 cases per year to 3.8% for a reader examining 15,000 cases per year. CDR did not change with reader volume (OR 1.0007, 95%CI 0.993 to 1.002, p=0.3). On average 7.55 cancers were detected (by the first reader) per 1000 cases read. Discrepant rate decreased with increasing volume of cases read (OR 0.981, 95%CI 0.976 to 0.987, p<0.001), from 0.97 for a reader examining 2000 cases per year to 0.46 for a reader examining 15,000 cases per year, see supplementary Figure 1 and Table 1. Discrepant rate is expressed as number of discrepant cases per 1000 cases read.

Performance between different occupational groups

There were no differences between consultant radiologists and radiography advanced practitioners for recall rate (OR 0.9985 CI 0.95, 1.05) or CDR (OR 0.9761 CI 0.95, 1.00) or PPV (OR 0.9750 CI 0.92, 1.03). Consultant radiographers had significantly higher recall rate (OR 1.105, CI 1.02, 1.19, p-value= 0.012) and lower PPV (OR 0.874, CI 0.80, 0.95, p=0.002) compared to consultant radiologists, but CDR did not differ (OR 0.9763, CI 0.94, 1.02, p=0.238). Interaction terms between reader role and volume were not statistically significant and not further included in the models.

Discussion

This study has examined the effect of reader volume on performance measures in the NHSBSP over a 3 year period, including 6.1 million reads by 759 readers. We found that PPV increased steadily with volume read from 12.9% for readers examining 2000 screening mammograms, through 14.4% for those reading 5000 screening mammograms, to 19.9% for readers examining 15,000 screening mammograms . This was driven by decreases in RR and no change in CDR. There was an association between increases in reader volume and a decrease in discrepant rate but interpretation is difficult because there are biases including reader pairing and the arbitration system.

A problem with available research is that the majority has not been based on actual screening data but rather assessing performance in test sets and correlating it with a readers reported reader volume. In 2005 Moss et al reviewed 5 studies where accuracy of reporting was related to reading volume. Three of the studies suggested a positive association between reading volume and sensitivity. Of the two studies showing no association one was underpowered in terms of a low number of film readers and the other included very low volume film readers [14]. Evidence from test set type studies in the USA and Australia have concluded that specificity is improved with higher initial experience, higher annual interpretation volume and access to regular feedback but that associations with sensitivity are more difficult to identify [15,16, 17]. Rawashdeh et al also found that for low volume readers (less than 1000 reads/annum) their performance cannot be compensated for by years of experience [17]. Theberge et al assessed the relation of volume to accuracy in the Quebec Breast Cancer Screening Program. Radiologists interpreting at least 4000 mammograms annually experienced a 32% increase in accuracy (adjusted accuracy ratio = 1.32; 95% CI = 1.13 to 1.54) compared with those interpreting 500 to 999 mammograms annually. This increase in accuracy was attributed to a reduction in false-positive rate as total volume increased whilst sensitivity changed little with total volume (8).

Our findings are in keeping with those from Hoff et al who published national data on the influence of mammography reader volume on radiologists' performance from Breast Screen Norway [11]. They looked at 2,373,433 reads by 121 radiologists over a 10 year period covering the transition from analogue to digital and suggested annual reading volumes between 4000 and 10,000 were optimal with the main effect on specificity with fewer false positives at higher volumes. Our results are also similar to a previous small UK study looking at regional data in the East Midlands which demonstrated lower recall rates in high volume readers [9] although Duncan et al who looked at data in Scotland for the same time period found no differences in any of the observed performance measures for different reader volumes [10].

We were not able to correct for reader experience in our comparison of occupational groups, which may have confounded the analysis. Whilst advanced practitioners have been reporting screening mammograms within the NHSBSP for approximately 20 years, there still may remain systematic differences in experience between the groups. The effect on discrepant rate was not included because of bias. However CDR was stable across all groups. Advanced practitioners did not differ in the other 2 indicators. Consultant radiographers had significantly higher recall rate and lower PPV compared to radiologists. Previous research in this area is limited but several studies have found that radiographer film readers have lower specificity but equivalent sensitivity to radiologists. A small meta-analysis in 2008 concluded that in a screening setting radiographers scored higher false positive rates with similar sensitivities compared with radiologists [18]. In 2012 Bennett et al published results of an observational study in the UK examining the performance of screening units in which a proportion of mammograms were double read using "non-discordant radiographer only (double) reading". The results showed an increase in recall rates in the pilot units relative to the comparison units at both prevalent and incident screens . There was no evidence to suggest a difference in CDR between the pilot units and the comparison units [19]. It is difficult to explain why our data has shown consultant radiographers had significantly higher recall rates compared to radiologists whilst advanced practitioners did not. It maybe that the additional responsibility assigned to consultant radiographers is associated with increased caution but that is only supposition.

This study is part of a growing evidence base that increased reader volume is associated with improved PPV, at least up to examining 10-15 thousand women's mammograms per year. The reason for this effect is unclear. The experience of examining a large volume of mammograms may contribute to improved accuracy. However, causation may be in the opposite direction, with readers with better PPV being encouraged to take on greater reading volume. A third explanation is that readers who examine increased volumes necessarily have longer reading sessions, and there is reasonable evidence that specificity improves with time on task and batch reading [20,21].

Interpreting these results for policy-making is complex. Whilst there is an overall association between reader volume and PPV, there is very large inter-reader variability. Our results suggest that the minimum reading volumes in Europe rather than the US may be more effective where practicable. Optimisation of reader pairings in breast units may be feasible with pairing of low and high volume readers. Brennan et al have examined this concept using test sets and concluded that strategic pairings may maximise the benefits of double reading [22].However there are significant practical issues to this approach especially in units where there are staff shortages. For population based screening programmes to be successful cancer detection rates need to be optimised whilst keeping recall rates as low as possible to avoid unnecessary recalls for false positive cases. It is important to consider not only how many cancers were detected at screening, but the types of cancer detected and the associated benefit and harm to the women screened. Recent large UK studies provide conflicting evidence on this. A recent analysis of 11.3 million screening tests in the English NHS breast screening programme showed that increasing recall rates above a threshold (7% for prevalent round and 4% for incident round) was almost exclusively associated with false positive recalls and a small increase in detection of low/intermediate grade ductal carcinoma in situ [23]. Another UK study of over 5 million screening episodes found higher recall rates were associated with lower interval cancer rates, with approximately one fewer interval cancer for every additional 80-84 recalls [24].

A major strength of our study is the large amount of real life data analysed from a short period of time through which imaging methods were stable. There was extremely low missing data due to the automated data processes, and robust quality assurance procedures. We found associations between reading volume and outcomes, but we did not correct for confounding factors and cannot make causal inferences in this cross-sectional study design. A further limitation is interpretation, particularly of discrepant rates, is challenging due to the biases from the hierarchical structures at screening centres influencing who is recalled for further tests. Whilst there is a large volume of data there remains statistical uncertainty at lower reading volumes. Furthermore readers who examined fewer cases would have a low cancer event rate resulting in very large confidence intervals. Therefore our results are not generalizable beyond the range of 1000-45000 cases per year. We have been unable to assess the effect of experience, training and access to feedback as this information is not readily available at the individual level. A high percentage of readers (42%) in our study did not appear to achieve the standard 5000 mammograms a year. However this is likely an artefact of our inclusion criteria, because 1000 of the 5000 requirement can be symptomatic mammograms which were excluded from our analysis.

In conclusion positive predictive value of screening increases with total volume of cases examined per reader through decreases in numbers of cases recalled with no concurrent change in numbers of cancers detected. CDR was stable across all occupational groups. There was some variation in recall rate and PPV across occupational groups, but we were not able to correct for reader experience, which may have confounded the analysis.

References

1. Mammography Quality Standards Act Regulations. U.S. Food & Drug Administration. https://www.fda.gov/Radiation-EmittingProducts/MammographyQualityStandardsActandProgram/Regulations/ucm110906.htm. Updated November 29, 2017. Accessed June 2020

2. BreastScreen Australia, "National accreditation standards," 2015 http://cancerscreening.gov.au/internet/screening/publishing.nsf/Content/br-accreditation/$File/standards.pdf. Accessed June 2020

3. Perry N, Broeders MJ, de Wolf C, Törnberg S, Holland R, von Karsa L.(2006) European guidelines for quality assurance in breast cancer screening and diagnosis. Fourth edition. Luxembourg: Office for Official Publications of the European Communities.

4. Quality Assurance Guidelines for Breast Cancer Screening Radiology. (2005) NHS Cancer Screening Programmes (NHSBSP Publication No 59).

5.Gur D, Bandos AI, Cohen CS, et al. (2008) The 'Laboratory Effect': Comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretation. Radiology 249(1):47–53.

6. Buist DS, Anderson ML, Haneuse SJ, et al (2011) Influence of annual interpretive volume on screening mammography performance in the United States. Radiology 259(1):72–84.

7. Smith-Bindman R, Chu P, Miglioretti DL, et al (2005) Physician predictors of mammographic accuracy. J Natl Cancer Inst 97(5):358–367.

8. Théberge I, Chang SL, Vandal N, et al (2014) Radiologist interpretive volume and breast cancer screening accuracy in a Canadian organized screening program. J Natl Cancer Inst 106(3):djt461.

9. Cornford E, Reed J, Murphy A, Bennett R, Evans A (2011) Optimal screening mammography reading volumes; evidence from real life in the East Midlands region of the NHS Breast Screening Programme. Clin Radiol 66(2):103–107.

10. Duncan KA, Scott NW (2011) Is film-reading performance related to the number of films read? The Scottish experience. Clin Radiol 66(2):99–102.

11. Hoff SR, Myklebust TA, Lee CI, Hofvind S (2019) Influence of mammography volume on radiologists' performance: results from BreastScreen Norway. Radiology 292:289–296.

12. Department of Health (2000) *The NHS Cancer plan: a plan for investment, a plan for reform*. Department of Health, London. Available via https://webarchive.nationalarchives.gov.uk/20130222181549/http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_4014513.pdf. Accessed 22nd November 2020

13. Public Health England (2016),Guidance on who can undertake arbitration. Public Health England, London. Available via https://www.gov.uk/government/publications/breast-screening-arbitration-guidance.Accessed 22nd November 2020

14. Moss SM, Blanks RG, Bennett RL (2005) Is radiologists' volume of mammography reading related to accuracy? A critical review of the literature. ClinRadiol 60:623-626.

15. Elmore JG, Jackson SL, Abraham L, et al (2009) Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy. Radiology 253:641-651.

16. Buist DS, Anderson ML, Smith RA, et al (2014) Effect of Radiologists' Diagnostic Work-up Volume on Interpretive Performance. Radiology 273):351-364.

17. Rawashdeh M, Lee W, Bourne R, et al (2013) Markers of good performance in mammography depend on number of annual readings. Radiology 269:61–7.

18. van den Biggelaar FJHM, Nelemans PJ, Flobbe K (2008) Performance of radiographers in mammogram interpretation: a systematic review. Breast17:87–92.

19. Bennett RL, Sellars SJ, Blanks RG, Moss SM (2012) An observational study to evaluate the performance of units using two radiographers to read screening mammograms. Clin Radiol.67:114–121.

20. Taylor-Phillips S, Wallis MG, Jenkinson D, et al (2016). Effect of Using the Same vs Different Order for Second Readings of Screening Mammograms on Rates of Breast Cancer Detection. JAMA 315(18):1956-1965

21. Burnside ES, Park JM, Fine JP, Sisney GA (2005) The use of batch reading to improve the performance of screening mammography. AJR Am J Roentgenol.185 (3):790-6.

22. Brennan PC, Ganesan A, Eckstein MP, et al (2019) Benefits of independent double reading in digital mammography: a theoretical evaluation of all possible pairing methodologies. Acad Radiol.; 26(6): 717–723.

23. Blanks RG, Given-Wilson RM, Cohen SL, Patnick J, Alison RJ, Wallis MG. (2019) An analysis of 11.3 million screening tests examining the association between recall and cancer detection rates in the English NHS breast cancer screening programme Eur Radiol 29(7):3812-3819.

24. Burnside ES, Vulkan D, Blanks RG, Duffy SW (2018) Association between Screening Mammography Recall Rate and Interval Cancers in the UK Breast Cancer Service Screening Program: A Cohort Study. Radiology 288(1):47-54.