Radiology: Imaging Cancer

The relationship between mammography readers real-life performance and performance in a test-set based assessment scheme in a national breast screening programme.

Journal:	*Radiology: Imaging Cancer*
Manuscript ID	RYCAN-20-0016.R2
Manuscript Type:	Original Research
Manuscript Categorization Terms:	Breast < 4. AREAS/SYSTEMS, Screening < 7. METHODOLOGY

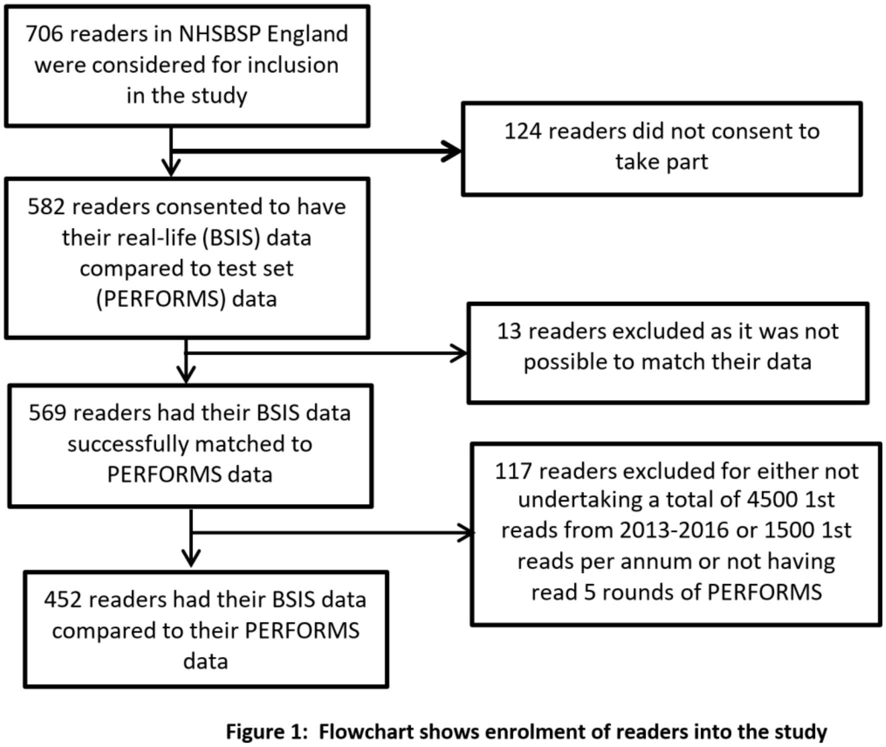

Figure 1: Flowchart shows enrollment of readers into the study.

397x322mm (115 x 115 DPI)

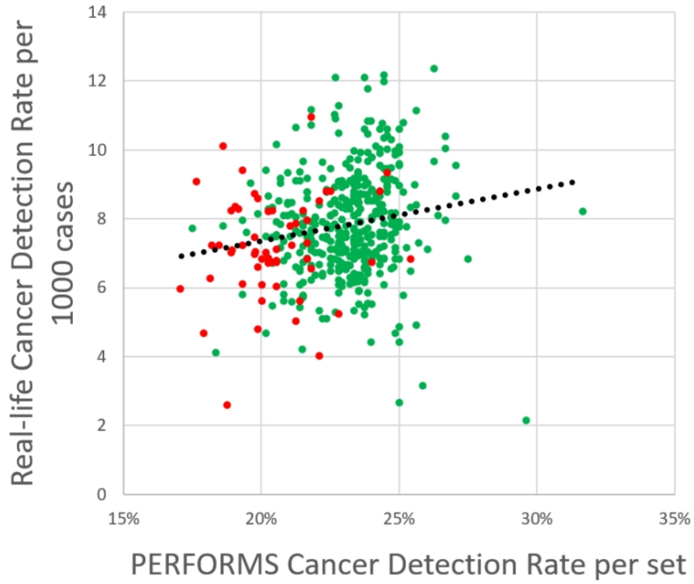

Figure 2A: Graph shows correlation between cancer detection in real life and in the PERFORMS test sets. Outliers are shown in red and non-outliers are shown in green.

Figure 2. Plots show correlation between (a) cancer detection rates, (b) recall rate, and (c) positive predictive value in real life and the PERFORMS tests sets.
Figure 2A: Graph shows correlation between cancer detection in real life and in the PERFORMS test sets.

273x192mm (115 x 115 DPI)

Figure 2B: Graph shows correlation between recall rate in real life and in PERFORMS test sets. Outliers are shown in red and non-outliers are shown in green.

Figure 2. Plots show correlation between (a) cancer detection rates, (b) recall rate, and (c) positive predictive value in real life and the PERFORMS tests sets.
Figure 2B: Graph shows correlation between recall rate in real life and in PERFORMS test sets.

289x197mm (115 x 115 DPI)

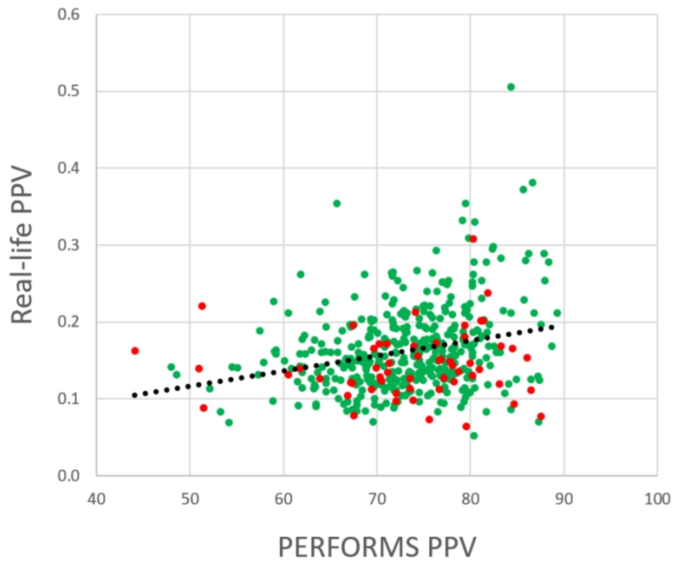

Figure 2C: Graph shows correlation between positive predictive value (PPV) in real life and PPV in the PERFORMS test sets. Note — PPV = positive predictive value. Outliers are shown in red and non-outliers are shown in green.

Figure 2. Plots show correlation between (a) cancer detection rates, (b) recall rate, and (c) positive predictive value in real life and the PERFORMS tests sets.
Figure 2C: Graph shows correlation between positive predictive value (PPV) in real life and PPV in the PERFORMS test sets.

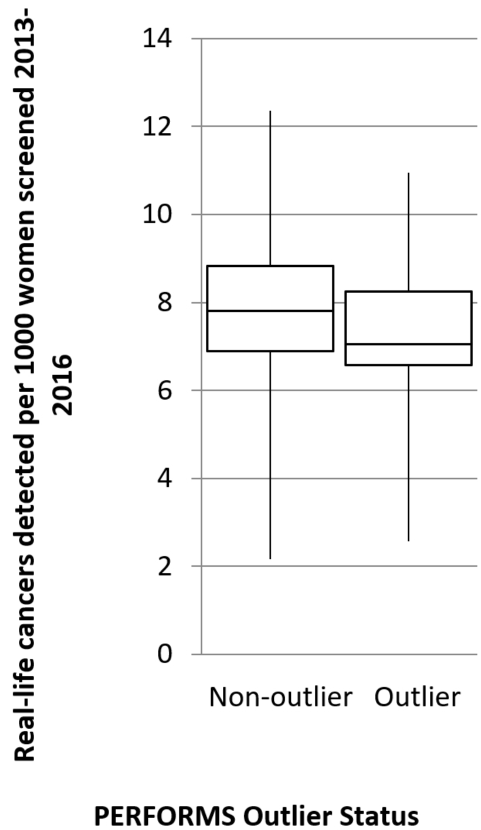

Figure 3a: Box-and-whisker plot shows real-life cancer detection rates based on whether or not readers were an "outlier" in the PERFORMS test sets.

Figure 3a

213x342mm (115 x 115 DPI)

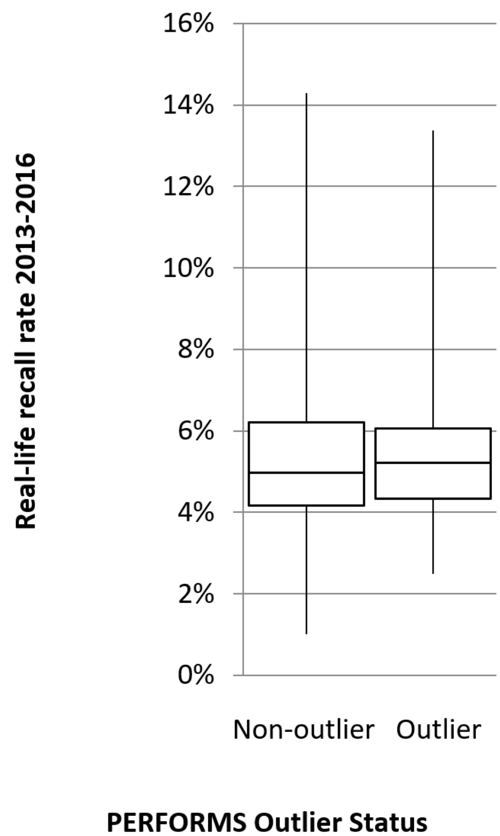

Figure 3b: Box-and-whisker plot shows real-life recall rates based on whether or not readers were an "outlier" in the PERFORMS test sets.

Figure 3b

219x347mm (115 x 115 DPI)

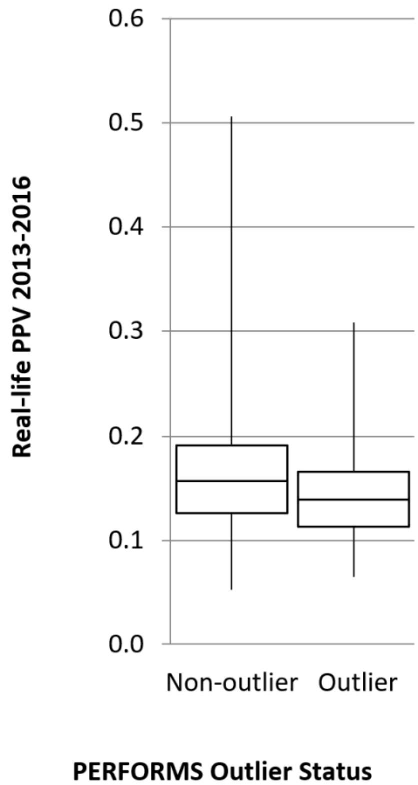

Figure 3c: Box-and-whisker plot shows real-life PPVs based on whether or not readers were an “outlier” in the PERFORMS test sets. Note — PPV = positive predictive value.

Figure 3c

221x371mm (115 x 115 DPI)

Relationship Between Mammography Readers Real-life Performance and Performance in a Test-set Based Assessment Scheme in a National Breast Screening Programme

Original Research

Abbreviations:
ANOVA = analysis of variance, BSIS = Breast Screening Information System, NHSBSP = National Health Service Breast Screening Programme, PACS = Picture Archiving and Communication System, PERFORMS = Personal Performance in Mammographic Screening, PPV = positive predictive value, ROC = receiver operating characteristic

Key Points:

-Readers' Breast cancer Screening Information System (BSIS) real-life performance significantly correlated with PERFORMS test for cancer detection rates (r = 0.179, P < .001), recall rates (r = 0.146, *P* = .002), and positive predictive value (r = 0.263, *P* < .001).

-Outliers in PERFORMS had significantly poorer real-life cancer detection rate and PPV of recall compared to the non-outlier group of readers.

-The PERFORMS tests has the potential to predict readers' performance and can be used to determine potential reading problems.

Summary statement:

The use of a test set based assessment scheme (PERFORMS) in a breast screening program has the potential to predict and identify poor performance in real-life.

Abstract

Purpose: To compare an individual's Personal Performance in Mammographic Screening (PERFORMS) score with their Breast Screening Information System (BSIS) real-life performance data and determine which parameters in the PERFORMS scheme offer the best reflection of BSIS real-life performance metrics.

Methods: In this retrospective study, the BSIS real-life performance metrics of individual readers (*n* = 452) in the NHS Breast Screening Programme (NHSBSP) in England were compared with performance in the test-set based assessment scheme over a 3-year period from 2013-2016. Cancer detection rate, recall rate, and positive predictive value (PPV) were calculated for each reader, for both real-life screening and the PERFORMS test. For each metric, real-life and test-set versions were compared using a Pearson correlation.

The real-life cancer detection rate, recall rate, and PPV of outliers were compared against other readers (non-outliers) using ANOVA.

Results: BSIS real-life cancer detection rates, recall rates, and PPV showed positive correlations with the equivalent PERFORMS measures ($P < 0.001$, $P = 0.002$, $P < 0.001$, respectively). The mean real-life cancer detection rate (CDR) of PERFORMS outliers was 7.2 per 1000 women screened and was significantly lower than other readers (non-outliers) where the real-life cancer detection rate was 7.9 ($P = 0.002$). The mean real-life screening PPV of PERFORMS outliers was 0.14% and was significantly lower than the non-outlier group who had a mean PPV of 0.17% ($P = 0.006$).

Conclusions: The use of test-set based assessment schemes in a breast screening program has the potential to predict and identify poor performance in real-life.

820 Jorie Blvd., Suite 200, Oak Brook, IL, 60523, 630-481-1071, RYimagingcancer@rsna.org

Introduction

There has been considerable interest in recent years for the assessment of performance of healthcare personnel. Individuals providing care have a duty to demonstrate satisfactory performance, forming part of appraisal and revalidation. Measuring individual performance has the potential to improve the quality of services offered, inform the public, determine potential problems, and provide supportive further training (1).

Breast radiology in the United Kingdom (UK), particularly in the context of the National Health Service Breast Screening Programme (NHSBSP), has always had its performance heavily audited as part of the quality assurance process which is integral to the service. Data on each of the screening centers has been collected and published since programme inception in 1988 (2). In addition, to provide a measure of individual performance, a test set based system called PERFORMS (Personal Performance in Mammographic Screening) has been running for over 30 years (3). Participants whose performance in the scheme is below a minimum acceptable standard (statistically significantly lower than that of the main body of readers) are flagged up as ‘outliers’ and further action is taken, such as reviewing practice, offering suggestions, or further training.

There has been criticism that test-set based performance schemes may suffer from a “laboratory effect” and not be a true reflection of real-life performance. Many studies demonstrate that experimental conditions can affect human behaviour (4). Test sets, by their very nature, are heavily enriched with cancer cases and the reader knows that any decisions they make in the test environment will have no patient impact and so reading behaviour may be altered (5).

Recently, the UK Breast Screening Information System (BSIS), which provides national and local performance statistics for the NHSBSP, has produced individual real-life performance data over rolling three-year periods. The aim of this study is to compare an individual’s PERFORMS test set scores with their real-life performance data and determine which parameters in the PERFORMS scheme offer the best reflection of real-life performance metrics. In addition, this study aims to determine whether the 'outlier' status in the PERFORMS scheme is a true predictor of poor performance in real life.

Materials and Methods

Study Design

All 706 readers who interpret screening mammograms for the NHSBSP in England and who take part in the PERFORMS self-assessment test were invited to participate in the study. Ethics approval was waived, following discussion with the local Research and Development Team as this retrospective comparison was considered to represent an audit of current practice. The study was carried out in accordance with the local Information System Security Policy, Data Protection Policy, and associated Codes of Practice and Guidelines, with participants giving informed consent for their performance data to be accessed.

A total of 582 readers consented for their real-life data to be accessed for the study. Real-life data were obtained from BSIS for the three-year period 2013-2016. Study participants had to have completed at least five rounds of the PERFORMS self-assessment scheme (i.e. 5 sets of 60 cases) within 36 months of the BSIS real-life screening data period. The NHSBSP requires readers to interpret 5000 mammograms each year, but at least 1500 of these have to be as a first reader (3). Consequently, participants had to read at least 1500 screening cases per year as a first reader, and no less than a total of 4500 cases as a first reader over the three-year period of the study to be included. In additional, participants were excluded if their real-life data could not be identified or matched with their PERFORMS data. Consequently, a total of 452 readers were available for the comparison. The flow chart in Figure 1, outlines the recruitment process and exclusion criteria.

PERFORMS Image Assessment

The PERFORMS scheme involves the circulation of test sets of 60 challenging cases, consisting of normal, benign, and abnormal mammograms. The test sets are heavily enriched with biopsy proven cancers (typically around 35%), with radiological features of masses, calcifications, asymmetries and distortions. Benign and normal cases are either biopsy proven or have at least three years of mammographic follow-up. Cases are chosen by the scheme organisers in conjunction with a national panel of *ten* expert breast radiologists with more than 20 years of experience working in the NHSBSP from a pool contributed by all UK screening centres. PERFORMS is currently undertaken by over 800 readers in the UK (6) as part of the quality assurance for the NHSBSP (7). Readers in the UK screening program include board certified radiologists, radiographers, or breast clinicians (doctors who are not radiologists working in the field of breast diagnosis). Non-radiologists typically make up half the readers in the UK programme and are trained to Masters level or equivalent and, along with the radiologists, have to undertake the reading of a minimum of 5000 mammograms per year (8).

The test-set images are uploaded to the Picture Archiving and Communication System (PACS) at each screening centre where they can be viewed. Readers' findings are recorded on a password protected website and participants receive immediate feedback on each case at the end of the set, compared to pathology and an opinion derived from a national panel of experts, who provide a commentary on the radiological appearances of the cancers and the appropriateness of recall for the normal and benign cases. Once completed by all readers, comprehensive performance statistics are produced providing an individual with a comparison with their peers nationally. Data is produced on correct recall for further assessment, correct return to normal screening, cancer detection rate, and the positive and negative predictive value of recall based on pathology.

Test Standards

The NHSBSP uses double reading as standard and so the performance data produced primarily focuses on the opinion of the individual as a first reader. In many centers, the second reader is not blinded to the opinion of the first reader and so the first read is the only truly unbiased read. The data extracted included a unique reader code, screening center name, number of cases read as first reader, number of recalled cases, cancers detected as first reader, as well as rate of discrepant cancers per year (defined as cancers missed by the first reader that where subsequently identified by the second reader). Comparative results from the PERFORMS tests sets were obtained from the PERFORMS data base which consisted of reader ID, screening center name, correct and incorrect recall, correct return to screening, and missed cancer rates.

Measures of sensitivity were selected to be analogous in real-life screening and in test-set based performance. In real-life screening, the cancer detection rate was calculated as the number of women in whom cancer was detected per 1000 women screened. For PERFORMS, the cancer detection rate was calculated as the percentage of cancers detected out of the total number of cases in the test set. Positive predictive value was calculated as the total number of cancers detected out of the total number of cases recalled, for both real-life screening performance and the test-set based performance; the number of "true positives" divided by the number of "true positives" plus "false positives". The real-life BSIS data cannot provide a true specificity measure or a negative predictive value (NPV). Due to the development of cancers between screening rounds (interval cancers), determining which cases are true and false negatives will not become apparent for many years. Consequently, in real-life screening the recall rate is used as a proxy for specificity. Recall rate was calculated as the total number of cases recalled out of the total number of cases read, for both the real-life screening and test-set based performance measures.

Statistical Analysis

Cancer detection rate, recall rate, and positive predictive value (PPV) measures were calculated from the PERFORMS data and from the BSIS real-life data, yielding two values per reader for each metric: one real-life screening-based value and one test-set based value. For each of these measures, a Pearson correlation between the PERFORMS test-set data and BSIS real-life screening data was examined. Further analysis assessed whether those readers whose performance on the PERFORMS test was deemed to be below the minimum acceptable standard (the outliers) had significantly poorer performance on the BSIS real-life screening measures. PERFORMS outliers are readers whose test performance falls more than one and a half times the inter-quartile range below the 25th percentile in terms of either cancer detection rate in the PERFORMS test set or the area under the curve of the receiver operating characteristic (ROC) analysis of their test set performance (or both). For the purposes of this study, any reader who had been an outlier on any of the PERFORMS test-sets included in three-year period, were allocated into an 'Outliers' group. The real-life cancer detection rates, recall rates, and PPVs of PERFORMS outliers were then compared against those of other readers using analysis of variance (ANOVA). The α-level for statistical significance was set at .05 for all analyses. Statistical calculations were performed using the IBM SPSS Statistics (version 23.0) statistical software (SPSS Inc., Chicago, IL).

Results

Participant Performance Overview

In total, 452 participants (238 board certified radiologists, 193 radiographer readers, and 21 breast clinicians) consented and were eligible to take part in the study. The mean cancer detection rate from the BSIS real-life data was 7.79 per 1000 women screened (0.78%) with a mean recall rate of 5.29%. Each PERFORMS test set of 60 cases is heavily enriched with cancers; the number of cancer cases varied between 34 and 38 for the PERFORMS sets included in this study. The mean cancer detection rate in the PERFORMS test sets was 22.86% with a mean recall rate of 37.49%. A summary of the BSIS real-life and PERFORMS performance measures for the participants is given in Table 1.

Test Measures Assessed from BSIS Real-life and PERFORMS Correlate

BSIS real-life cancer detection rates, recall rates, and PPVs showed significant positive correlations with the equivalent PERFORMS measures (n = 452). Readers with a higher cancer detection rate in real-life tended to have a higher cancer detection rate in PERFORMS (Pearson's Correlation: r = 0.179, P < .001, two tails; Figure 2A). Readers with a higher recall rate in real-life screening tended to have a higher recall rate in PERFORMS (Pearson's Correlation: r = 0.146, P = .002, two tails; Figure 2B). PPV, the probability that a patient recalled following screening mammography has a confirmed breast malignancy, reflects a combination of cancer detection rate and recall rate. Readers with a higher PPV in real-life screening tended to have a higher PPV in PERFORMS (Pearson's Correlation: r = 0.263, P < .001, two tails; Figure 2C). It is noted that, as PPV is affected by the prevalence of the disease, PPV in the test-set data was considerably higher than in the real-life data, reflecting the difference in the prevalence of cancers in the two data-sets.

Comparison of Outliers and Nonoutliers

Outliers in the PERFORMS scheme were found to have significantly lower performance than other readers in real-life screening in terms of cancer detection rate and PPV, but did not differ significantly in terms of recall rate (Table 2). The mean BSIS real-life screening cancer detection rate of PERFORMS outliers was 7.2 per 1000 women screened and was significantly lower than other readers (non-outliers) where the cancer detection rate was 7.9 per 1000 women screened (ANOVA $F(1, 450) = 9.78$, p = .002, ω = .014) (Figure 3A). The mean BSIS real-life screening recall rate of PERFORMS outliers was 5.5% and was not different from that of other readers who had a mean of

5·3% (ANOVA F(1, 450) = 0.67, P = .415, ω = .003) (Figure 3B). The mean BSIS real-life screening PPV of PERFORMS outliers was 0.14% and was significantly lower than the non-outlier group who had a mean PPV of 0.17% (ANOVA F(1, 450) = 7.75, P = .006, ω = .012) (Figure 3C).

Discussion

This study was designed to determine if performance in the PERFORMS test set scheme reflected BSIS real-life performance. Test set performance demonstrated significant positive correlations with the BSIS real-life performance metrics produced by the UK screening programme, i.e. cancer detection rate(r = 0.179, P < .001), recall rate(r = 0.146, P = .002), PPV(r = 0.263, P < .001) all showed strong correlations For breast cancer screening to be successful, cancer detection rates need to be optimized, but at the same time recall rates need to be kept as low as possible to avoid false positive interpretation and recalls. There will always be a trade-off between recalling women for further investigation and detecting cancers, which is reflected in the PPV. Recall rates act as a proxy for specificity in real-life screening, due to the difficulty in identifying true negatives and false negatives at the time of reading. However, recall rates are not a perfect measure of specificity. Recall rates need to be interpreted in conjunction with cancer detection – both low and high recall rates would be acceptable in the context of high cancer detection, whereas in isolation extreme recall rates may raise concerns about a reader's performance.

Correlation between BSIS real-life recall rates and PERFORMS correct recall rates was the least strong of the performance metrics, although it did reach statistical significance (r = 0.146, P = .002). One of the criticisms of test sets is that reading behaviour may be altered. This weaker correlation is probably not surprising as it has previously demonstrated that recall rates are particularly prone to this 'laboratory' effect, as readers know that flagging a patient for recall will have no impact on patient care (4).

Previous studies comparing test-set and real-life performance have shown consistently positive relationships, albeit weak in some instances (9-11). One of the strengths of this study is that is has been possible to compare real-life performance data with results from a test-set scheme in a large group of readers. Soh et al reported reasonable levels (*P*<.01) of agreement between actual clinical reporting and test set conditions, although increased sensitivity was seen under test set conditions (11). This study of 452 participants demonstrated much stronger associations than a previous smaller study of 40 readers from one UK region taking part in the same PERFORMS scheme in 2005 and 2006 (10). PPV of recall demonstrated the strongest correlation between BSIS real-life and PERFORMS data for all participants. PPV is one of the most useful measures of performance (12).

Real-life performance data is often considered the reference standard. However, the accuracy of sensitivity and specificity of real-life breast cancer screening data is problematic (13). Reader sensitivity, which is defined as the proportion of patients with breast cancer reported as positive, is not known for several years until interval cancer data becomes available and even then real life data may not be updated to reflect this. Due to this unavoidable time lag, the opportunity to introduce timely interventions to improve performance is lost. Similarly, when measuring specificity as the proportion of disease-free patients reported as negative, a truly negative mammogram will not be apparent until after the next screening round at the earliest. One of the advantages of test sets like PERFORMS is that normal, benign, and malignant cases with known, biopsy proven outcomes and appropriate follow up can be selected for inclusion, providing potentially more accurate performance metrics. For instance, when choosing cases for PERFORMS, a normal case will only be included if the mammogram at the next screening round three years later is also normal.

One of the key functions of measuring performance is to identify potential problems at the earliest opportunity to allow interventions to change practice. Real life data is by its very nature retrospective. Cancer detection rates of around 7-8 per 1000 women screened mean that an individual reader is exposed to relatively few cancers each year. Consequently, it can be difficult to identify poor performance because of the statistical instability from the relatively small number of cancer cases, similar problems are encountered when measuring performance in NHSBSP screening centres with the smallest number of clients (14). BSIS audit data are combined over a three-year period to improve the statistical robustness of the performance measures, but even so many years of poor performance may occur before this becomes apparent through clinical audit, resulting in potential harm to the screening population. For many years the PERFORMS scheme has flagged up poor performance outliers where metrics have deviated significantly from the mean. Individuals and the regional quality assurance office are notified so that corrective measured can be instigated such as reviewing practice or further training. PERFORMS has the potential to identify under performance at a much earlier stage than real-life data, perhaps even before a reader takes part in the screening programme as part of an end of training or pre-employment assessment. If test sets are to be used in this way, then it is crucial that the results are validated against real-life data. In this study, being a poor performance outlier in PERFORMS was able to predict poor real-life performance with outliers have significantly poorer real-life cancer detection rate and PPV of recall compared to the non-outlier group of readThis study does have limitations. Nearly 20% of PERFORMS participants (124 readers) declined to have their data used and so this has to be considered a potential source of bias. Further work is needed to understand if this group had any particular characteristics.

In conclusion, there are significant correlations between real-life readers' performance in a breast screening programme and their performance on metrics generated from a test-set based assessment scheme such as PERFORMS. Readers' positive predictive value of recall in real-life screening and the test-sets showed the strongest correlations. The use of test-set based assessment schemes has the potential to predict and identify potential poor performance outliers in real-life screening, enabling corrective measures to be implemented in a timely fashion.

Funding:.

References

1. Hall LH, Johnson J, Watt I, Tsipa A, O'Connor DB. Healthcare staff wellbeing, burnout and patient safety: a systematic review. *PLoS One* 2016; 11(7):e0159015. **doi:** 10.1371/journal.pone.0159015. Published July 8, 2016. Accessed October 10, 2019.

2. Cohen SL, Blanks RG, Jenkins J, Kearins O. Role of performance metrics in breast cancer screening imaging – where are we and where should we be? *Clin Radiol* 2018; 73:381–88. doi: 10.1016/j.crad.2017.12.012

3. Chen Y, Gale A. Performance assessment using standardized data sets: The PERFORMS scheme in breast screening and other domains. In: Samei E, Krupinski EA, eds. The Handbook of Medical Image Perception and Techniques. 2nd ed. Cambridge: Cambridge University Press, 2018; 328–42.

4. Egglin TKP, Feinstein AR. Context bias: a problem in diagnostic radiology. *JAMA* 1996; 276(21):1752–55. doi: 10.1001/jama.1996.03540210060035

5. Gur D, Bandos AI, Cohen CS, Hakim CM, Hardesty LA, Ganott MA, et al. The 'Laboratory Effect': Comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretation. *Radiology* 2008; 249(1):47–53. doi: 10.1148/radiol.2491072025

6. NHS public health functions agreement 2018-19; Public health functions to be exercised by NHS England. EU, International and Prevention Programmes, Global and Public Health Group, Public Health Systems and Strategy; Department of Health and Social Care website. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/694130/nhs-public-functions-agreement-2018-2019.pdf . Published March 26, 2018. Accessed May 12, 2018.

7. Programme Specific Operating Model for Quality Assurance of breast screening Programmes. Government UK website. https://www.gov.uk/government/publications/breast-screening-programme-specific-operating-model Published July 2017 . Accessed August 15, 2019

8. Quality assurance guidelines for breast cancer screening radiology. NHS Cancer Screening Programmes website. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/764452/Quality_assurance_guidelines_for_breast_cancer_screening_radiology_updated_Dec_2018.pdf. Published March 2011. Accessed April 6, 2020

9. Soh BP, Lee W, Kench PL, Reed WM, McEntee MF, Poulos A, et al. Assessing reader performance in radiology, an imperfect science: Lessons from breast screening. *Clin Radiol* 2012; 67(7):623–28. doi: 10.1016/j.crad.2012.02.007

10. Scott HJ, Evans A, Gale AG, Murphy A, Reed J. The relationship between real life breast screening and an annual self-assessment scheme. Proceedings of SPIE, Medical Imaging 2009: Image Perception, Observer Performance, and Technology Assessment; 72631E. **doi:** 10.1117/12.811003. Published March 12, 2009. Accessed September 22, 2019

11. Soh BP, Lee W, McEntee MF, Kench PL, Warren M., Heard R, et al. Screening Mammography: Test Set data can reasonably describe actual clinical reporting. *Radiology* 2013; 268:46–53. doi: 10.1148/radiol.13122399

12. Bennett RL, Blanks RG. Should a standard be defined for the Positive Predictive Value (PPV) of recall in the UK NHS Breast Screening Programme? Breast J 2007; 16(1):55–9. doi: 10.1016/j.breast.2006.05.008

13. Rutter CM, Taplin S. Assessing mammographers' accuracy: a comparison of clinical and test performance. *J Clin Epidemiol* 2000; 53(5):443-450. doi: 10.1016/S0895-4356(99)00218-8

14. Blanks RG, Bennett RL, Wallis MG, Moss SM. Does individual programme size affect screening performance? Results from the United Kingdom NHS breast screening programme. *J Med Screen* 2002; 9:11–4. doi: 10.1136/jms.9.1.11

Tables

Table 1: Summary of Real-life and PERFORMS Performance Measures

	Real-life			PERFORMS		
Values	PPV (%)	Cancers detected per 1000 women (*n*)	Recall rate (%)	PPV (%)	Cancer detection rate (%)	Recall rate (%)
Mean ± standard deviation	0.16 ± 0.05	7.79 ± 1.55	5.29 ± 1.77	73.23 ± 7.29	22.86 ± 1.84	37.49 ± 5.86
95% confidence interval	0.16, 0.17	7.65, 7.93	5.12, 5.45	72.56, 73.91	22.69, 23.03	36.95, 38.03
Median (min, max)	0.15 (0.05, 0.51)	7.72 (2.16, 12.37)	5.02 (1.01, 14.29)	73.65 (44.12, 89.25)	23.06 (17.08, 31.67)	36.88 (23.89, 66.81)
25th and 75th percentile	0.13, 0.19	6.82, 8.76	4.17, 6.18	68.89, 78.17	21.67, 24.03	33.96, 40.28

Note — A total of 452 radiologists were assessed for real-life performance and PERFORMS. PPV = positive predictive value; CI = confidence interval; PERFORMS = Personal Performance in Mammographic Screening.

Table 2: Summary of Real-life and PERFORMS Performance Measures Based on Whether or not Readers were an "Outlier" in the PERFORMS Test Sets

	Real-life performance metrics					
	Number of cancers detected per 1000 women screened (*n*)		Recall rate (%)		Positive predictive value (%)	
	PERFORMS Outlier Status (2013-2016)					
Values	Non-outlier	Outlier	Non-outlier	Outlier	Non-outlier	Outlier
Mean ± standard deviation	7.9 ± 1.5	7.2 ± 1.5	5.3 ± 1.8	5.5 ± 1.8	0.17 ± 0.06	0.14 + 0.04
95% confidence interval	7.7, 8.0	6.8, 7.6	5.1, 5.4	5.0, 5.9	0.16, 0.17	0.13, 0.16
Median (min, max)	7.8 (2.2, 12.4)	7.1 (2.6, 11.0)	5.0 (1.0, 14.3)	5.2 (2.5, 13.4)	0.16 (0.05, 0.51)	0.14 (0.06, 0.31)
25th and 75th percentile	6.9, 8.8	6.6, 8.2	4.2, 6.2	4.3, 6.1	0.13, 0.19	0.11, 0.17
P value	0.002		0.415		0.006	

Note — There were a total of 396 non-outliers and 56 outliers. PPV = positive predictive value; CI = confidence interval; PERFORMS = Personal Performance in Mammographic Screening.

Figure 1: Flowchart shows enrolment of readers into the study.

Figure 2. Plots show correlation between **(a)** cancer detection rates, **(b)** recall rate, and **(c)** positive predictive value in real life and the PERFORMS tests sets.

Figure 2A: Graph shows correlation between cancer detection in real life and in the PERFORMS test sets.

Figure 2B: Graph shows correlation between recall rate in real life and in PERFORMS test sets.

Figure 2C: Graph shows correlation between positive predictive value (PPV) in real life and PPV in the PERFORMS test sets.

Figure 3: A total of 396 non-outliers and 56 outliers were assessed for their cancer detection rates per 1000 women, recall rates, and positive predictive value (PPV). Box-and-whisker plots show **(a)** real-life cancer detection rates, **(b)** real-life recall rates, and **(c)** real-life PPVs based on whether or not readers were an “outlier” in the PERFORMS test sets. The 95% confidence limits are shown on each plot.