

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

SCHOLARONE™  
Manuscripts

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

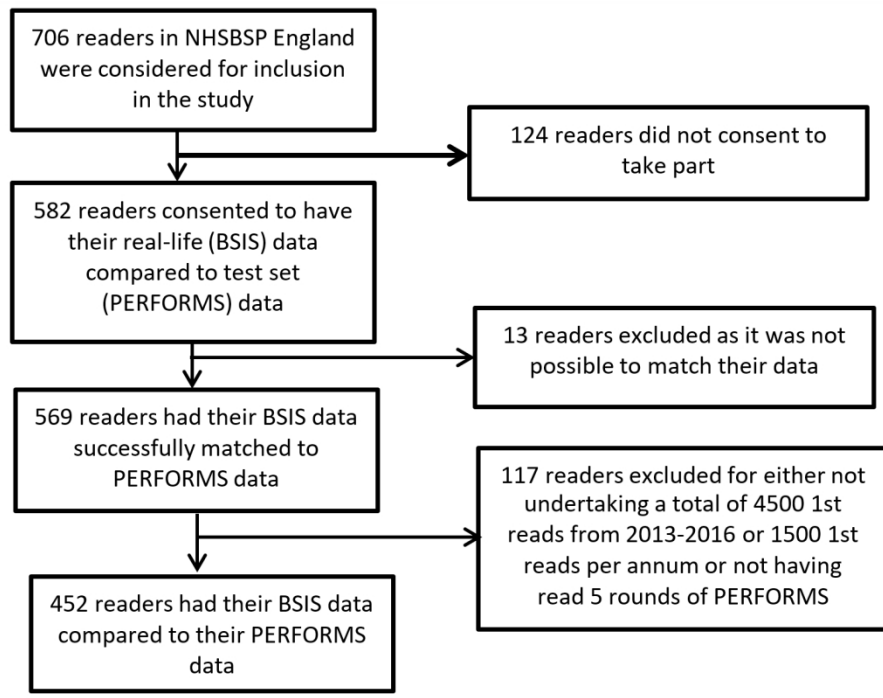
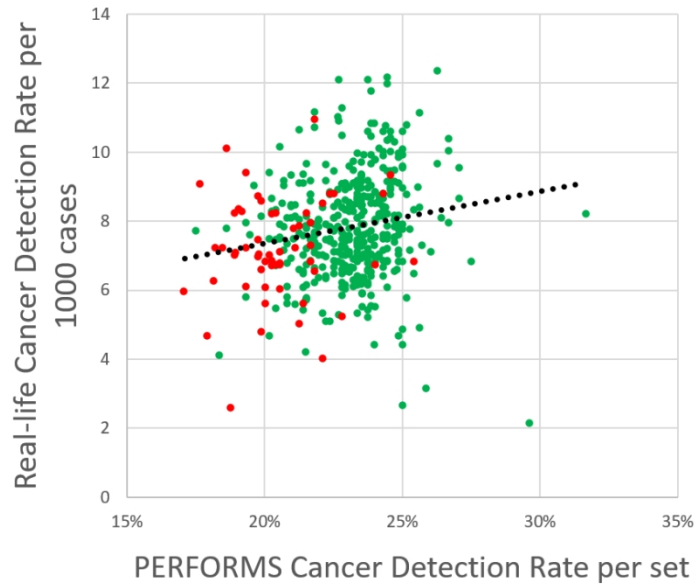


Figure 1: Flowchart shows enrolment of readers into the study

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

397x322mm (115 x 115 DPI)

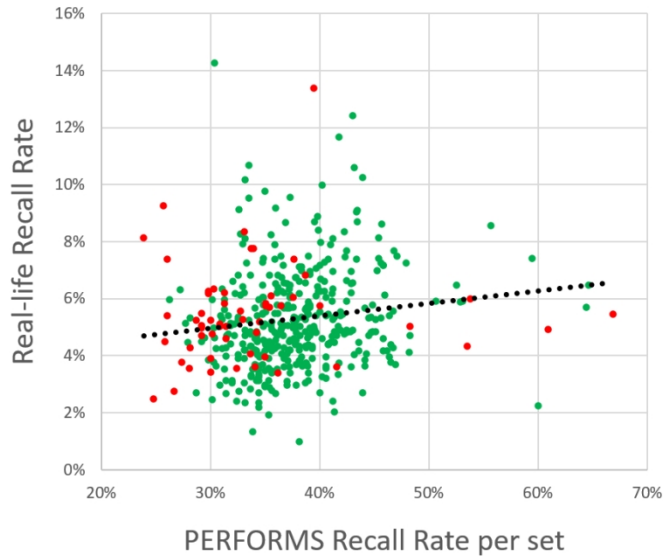


**Figure 2A:** Graph shows correlation between cancer detection in real life and in the PERFORMS test sets.  
Outliers are shown in red and non-outliers are shown in green.

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)

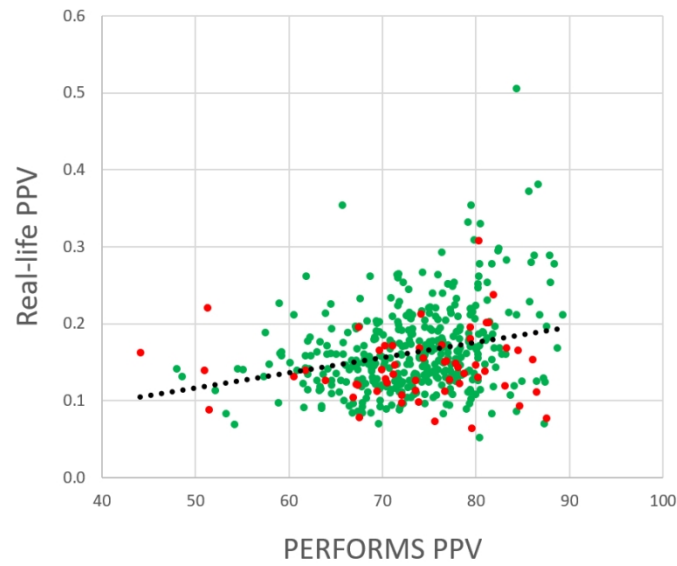


**Figure 2B:** Graph shows correlation between recall rate in real life and in PERFORMS test sets. Outliers are shown in red and non-outliers are shown in green.

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)

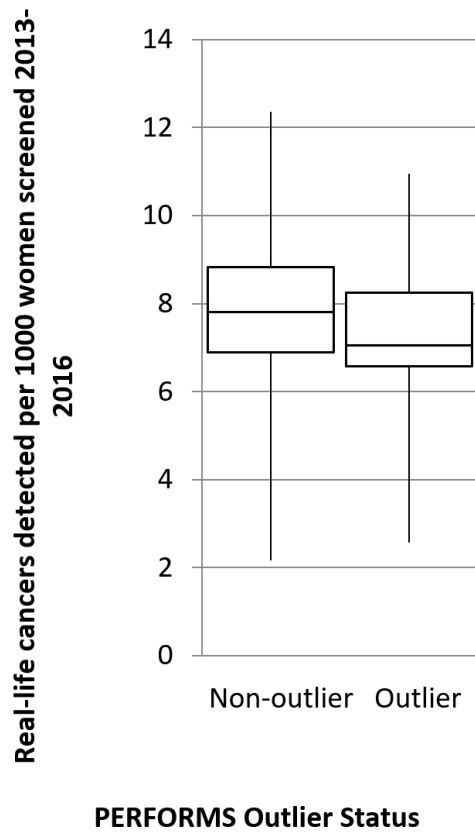


**Figure 2C:** Graph shows correlation between positive predictive value (PPV) in real life and PPV in the PERFORMS test sets. Note — PPV = positive predictive value. Outliers are shown in red and non-outliers are shown in green.

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.

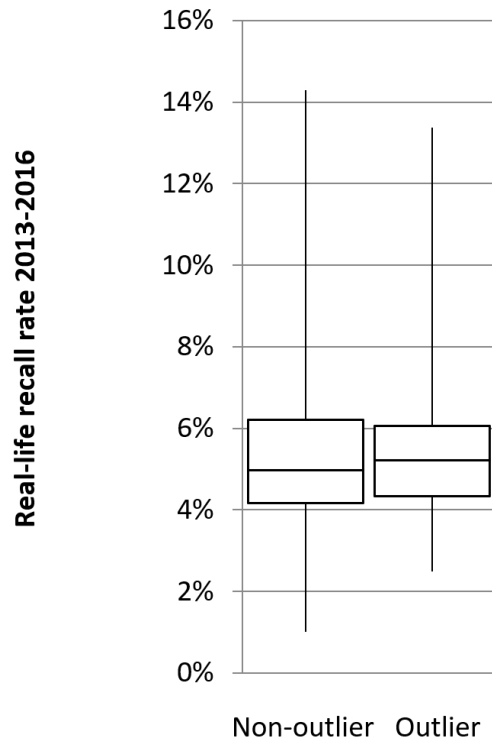
1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



**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)



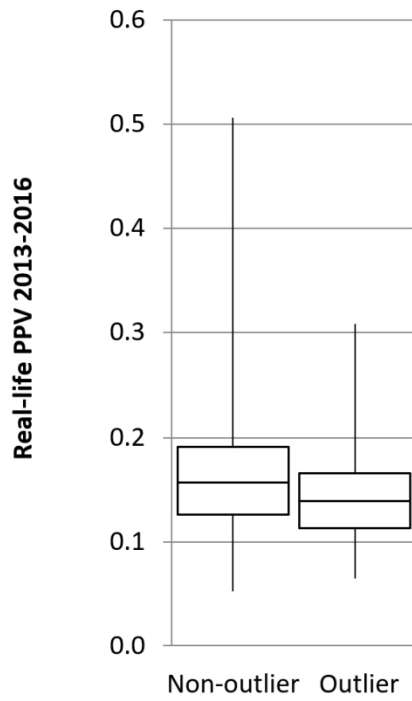
### PERFORMS Outlier Status

**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)

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



**PERFORMS Outlier Status**

**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)



1  
2  
3 1 **Relationship Between Mammography Readers Real-life Performance and Performance in a Test-**  
4 2 **set Based Assessment Scheme in a National Breast Screening Programme**  
5  
6  
7 3

8  
9 4 **Original Research**  
10  
11 5  
12

13 6 **Abbreviations:**

14 7 ANOVA = analysis of variance, BSIS = Breast Screening Information System, NHSBSP = National Health  
15 8 Service Breast Screening Programme, PACS = Picture Archiving and Communication System,  
16 9 PERFORMS = Personal Performance in Mammographic Screening, PPV = positive predictive value, ROC  
17 10 = receiver operating characteristic  
18  
19 11  
20 12

21 13 **Key Points:**

22  
23 14 -Readers' Breast cancer Screening Information System (BSIS) real-life performance significantly  
24 15 correlated with PERFORMS test for cancer detection rates ( $r = 0.179$ ,  $P < .001$ ), recall rates ( $r = 0.146$ ,  
25 16  $P = .002$ ), and positive predictive value ( $r = 0.263$ ,  $P < .001$ ).

26  
27  
28 17 -Outliers in PERFORMS had significantly poorer real-life cancer detection rate and PPV of recall  
29 18 compared to the non-outlier group of readers.

30  
31 19 -The PERFORMS tests has the potential to predict readers' performance and can be used to determine  
32 20 potential reading problems.  
33

34 21  
35 22 **Summary statement:**

36  
37  
38 23 The use of a test set based assessment scheme (PERFORMS) in a breast screening program has the  
39 24 potential to predict and identify poor performance in real-life.  
40  
41 25  
42  
43  
44 26  
45  
46  
47 27  
48  
49  
50 28  
51  
52  
53 29  
54  
55  
56 30  
57  
58  
59 31  
60

1  
2  
3 32 **Abstract**  
4  
5

6 33 **Purpose:** To compare an individual's Personal Performance in Mammographic Screening (PERFORMS)  
7  
8 34 score with their Breast Screening Information System (BSIS) real-life performance data and determine  
9  
10 35 which parameters in the PERFORMS scheme offer the best reflection of BSIS real-life performance  
11  
12  
13 36 metrics.

14  
15  
16 37 **Methods:** In this retrospective study, the BSIS real-life performance metrics of individual readers ( $n$   
17  
18 38 = 452) in the NHS Breast Screening Programme (NHSBSP) in England were compared with  
19  
20 39 performance in the test-set based assessment scheme over a 3-year period from 2013-2016. Cancer  
21  
22 40 detection rate, recall rate, and positive predictive value (PPV) were calculated for each reader, for  
23  
24 41 both real-life screening and the PERFORMS test. For each metric, real-life and test-set versions were  
25  
26 42 compared using a Pearson correlation.

27  
28  
29 43 The real-life cancer detection rate, recall rate, and PPV of outliers were compared against other  
30  
31 44 readers (non-outliers) using ANOVA.

32  
33  
34 45 **Results:** BSIS real-life cancer detection rates, recall rates, and PPV showed positive correlations with  
35  
36 46 the equivalent PERFORMS measures ( $P < 0.001$ ,  $P = 0.002$ ,  $P < 0.001$ , respectively). The mean real-life  
37  
38 47 cancer detection rate (CDR) of PERFORMS outliers was 7.2 per 1000 women screened and was  
39  
40 48 significantly lower than other readers (non-outliers) where the real-life cancer detection rate was 7.9  
41  
42 49 ( $P = 0.002$ ). The mean real-life screening PPV of PERFORMS outliers was 0.14% and was significantly  
43  
44 50 lower than the non-outlier group who had a mean PPV of 0.17% ( $P = 0.006$ ).

45  
46  
47  
48 51 **Conclusions:** The use of test-set based assessment schemes in a breast screening program has the  
49  
50 52 potential to predict and identify poor performance in real-life.  
51  
52

53  
54 53  
55

56  
57 54  
58  
59 55  
60

1  
2  
3 56 **Introduction**  
4  
5

6 57 There has been considerable interest in recent years for the assessment of performance of healthcare  
7  
8 58 personnel. Individuals providing care have a duty to demonstrate satisfactory performance, forming  
9  
10 59 part of appraisal and revalidation. Measuring individual performance has the potential to improve  
11  
12 60 the quality of services offered, inform the public, determine potential problems, and provide  
13  
14 61 supportive further training (1).  
15  
16  
17

18 62 Breast radiology in the United Kingdom (UK), particularly in the context of the National Health Service  
19  
20 63 Breast Screening Programme (NHSBSP), has always had its performance heavily audited as part of the  
21  
22 64 quality assurance process which is integral to the service. Data on each of the screening centers has  
23  
24 65 been collected and published since programme inception in 1988 (2). In addition, to provide a measure  
25  
26 66 of individual performance, a test set based system called PERFORMS (Personal Performance in  
27  
28 67 Mammographic Screening) has been running for over 30 years (3). Participants whose performance  
29  
30 68 in the scheme is below a minimum acceptable standard (statistically significantly lower than that of  
31  
32 69 the main body of readers) are flagged up as 'outliers' and further action is taken, such as reviewing  
33  
34 70 practice, offering suggestions, or further training.  
35  
36  
37  
38

39 71 There has been criticism that test-set based performance schemes may suffer from a "laboratory  
40  
41 72 effect" and not be a true reflection of real-life performance. Many studies demonstrate that  
42  
43 73 experimental conditions can affect human behaviour (4). Test sets, by their very nature, are heavily  
44  
45 74 enriched with cancer cases and the reader knows that any decisions they make in the test environment  
46  
47 75 will have no patient impact and so reading behaviour may be altered (5).  
48  
49  
50

51 76 Recently, the UK Breast Screening Information System (BSIS), which provides national and local  
52  
53 77 performance statistics for the NHSBSP, has produced individual real-life performance data over rolling  
54  
55 78 three-year periods. The aim of this study is to compare an individual's PERFORMS test set scores with  
56  
57 79 their real-life performance data and determine which parameters in the PERFORMS scheme offer the  
58  
59  
60

1  
2  
3 80 best reflection of real-life performance metrics. In addition, this study aims to determine whether the  
4  
5 81 'outlier' status in the PERFORMS scheme is a true predictor of poor performance in real life.  
6  
7

8 82  
9

## 10 11 83 **Materials and Methods**

### 12 13 14 84 *Study Design*

15  
16  
17 85 All 706 readers who interpret screening mammograms for the NHSBSP in England and who take part  
18  
19 86 in the PERFORMS self-assessment test were invited to participate in the study. Ethics approval was  
20  
21 87 waived, following discussion with the local Research and Development Team as this retrospective  
22  
23 88 comparison was considered to represent an audit of current practice. The study was carried out in  
24  
25 89 accordance with the local Information System Security Policy, Data Protection Policy, and associated  
26  
27 90 Codes of Practice and Guidelines, with participants giving informed consent for their performance  
28  
29 91 data to be accessed.  
30  
31  
32

33  
34 92 A total of 582 readers consented for their real-life data to be accessed for the study. Real-life data  
35  
36 93 were obtained from BSIS for the three-year period 2013-2016. Study participants had to have  
37  
38 94 completed at least five rounds of the PERFORMS self-assessment scheme (i.e. 5 sets of 60 cases)  
39  
40 95 within 36 months of the BSIS real-life screening data period. The NHSBSP requires readers to  
41  
42 96 interpret 5000 mammograms each year, but at least 1500 of these have to be as a first reader (3).  
43  
44 97 Consequently, participants had to read at least 1500 screening cases per year as a first reader, and  
45  
46 98 no less than a total of 4500 cases as a first reader over the three-year period of the study to be  
47  
48 99 included. In additional, participants were excluded if their real-life data could not be identified or  
49  
50 100 matched with their PERFORMS data. Consequently, a total of 452 readers were available for the  
51  
52 101 comparison. The flow chart in Figure 1, outlines the recruitment process and exclusion criteria.  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 103 *PERFORMS Image Assessment*  
4  
5

6 104 The PERFORMS scheme involves the circulation of test sets of 60 challenging cases, consisting of  
7  
8 105 normal, benign, and abnormal mammograms. The test sets are heavily enriched with biopsy proven  
9  
10 106 cancers (typically around 35%), with radiological features of masses, calcifications, asymmetries and  
11  
12 107 distortions. Benign and normal cases are either biopsy proven or have at least three years of  
13  
14 108 mammographic follow-up. Cases are chosen by the scheme organisers in conjunction with a  
15  
16 109 national panel of *ten* expert breast radiologists with more than 20 years of experience working in the  
17  
18 110 NHSBSP from a pool contributed by all UK screening centres. PERFORMS is currently undertaken by  
19  
20 111 over 800 readers in the UK (6) as part of the quality assurance for the NHSBSP (7). Readers in the UK  
21  
22 112 screening program include board certified radiologists, radiographers, or breast clinicians (doctors  
23  
24 113 who are not radiologists working in the field of breast diagnosis). Non-radiologists typically make up  
25  
26 114 half the readers in the UK programme and are trained to Masters level or equivalent and, along with  
27  
28 115 the radiologists, have to undertake the reading of a minimum of 5000 mammograms per year (8).  
29  
30  
31  
32  
33  
34 116 The test-set images are uploaded to the Picture Archiving and Communication System (PACS) at each  
35  
36 117 screening centre where they can be viewed. Readers' findings are recorded on a password  
37  
38 118 protected website and participants receive immediate feedback on each case at the end of the set,  
39  
40 119 compared to pathology and an opinion derived from a national panel of experts, who provide a  
41  
42 120 commentary on the radiological appearances of the cancers and the appropriateness of recall for the  
43  
44 121 normal and benign cases. Once completed by all readers, comprehensive performance statistics are  
45  
46 122 produced providing an individual with a comparison with their peers nationally. Data is produced on  
47  
48 123 correct recall for further assessment, correct return to normal screening, cancer detection rate, and  
49  
50 124 the positive and negative predictive value of recall based on pathology.  
51  
52  
53

54 125

55 126

56  
57  
58 127  
59  
60

1  
2  
3 128 *Test Standards*  
4  
5

6 129 The NHSBSP uses double reading as standard and so the performance data produced primarily  
7  
8 130 focuses on the opinion of the individual as a first reader. In many centers, the second reader is not  
9  
10 131 blinded to the opinion of the first reader and so the first read is the only truly unbiased read. The  
11  
12 132 data extracted included a unique reader code, screening center name, number of cases read as first  
13  
14 133 reader, number of recalled cases, cancers detected as first reader, as well as rate of discrepant  
15  
16 134 cancers per year (defined as cancers missed by the first reader that were subsequently identified  
17  
18 135 by the second reader). Comparative results from the PERFORMS tests sets were obtained from the  
19  
20 136 PERFORMS data base which consisted of reader ID, screening center name, correct and incorrect  
21  
22 137 recall, correct return to screening, and missed cancer rates.  
23  
24  
25

26  
27 138 Measures of sensitivity were selected to be analogous in real-life screening and in test-set based  
28  
29 139 performance. In real-life screening, the cancer detection rate was calculated as the number of  
30  
31 140 women in whom cancer was detected per 1000 women screened. For PERFORMS, the cancer  
32  
33 141 detection rate was calculated as the percentage of cancers detected out of the total number of cases  
34  
35 142 in the test set. Positive predictive value was calculated as the total number of cancers detected out  
36  
37 143 of the total number of cases recalled, for both real-life screening performance and the test-set based  
38  
39 144 performance; the number of "true positives" divided by the number of "true positives" plus "false  
40  
41 145 positives". The real-life BSIS data cannot provide a true specificity measure or a negative predictive  
42  
43 146 value (NPV). Due to the development of cancers between screening rounds (interval cancers),  
44  
45 147 determining which cases are true and false negatives will not become apparent for many years.  
46  
47 148 Consequently, in real-life screening the recall rate is used as a proxy for specificity. Recall rate was  
48  
49 149 calculated as the total number of cases recalled out of the total number of cases read, for both the  
50  
51 150 real-life screening and test-set based performance measures.  
52  
53  
54  
55

56 151  
57  
58

59 152  
60

1  
2  
3 153 *Statistical Analysis*  
4  
5

6 154 Cancer detection rate, recall rate, and positive predictive value (PPV) measures were calculated from  
7  
8 155 the PERFORMS data and from the BSIS real-life data, yielding two values per reader for each metric:  
9  
10 156 one real-life screening-based value and one test-set based value. For each of these measures, a  
11  
12 157 Pearson correlation between the PERFORMS test-set data and BSIS real-life screening data was  
13  
14 158 examined. Further analysis assessed whether those readers whose performance on the PERFORMS  
15  
16 159 test was deemed to be below the minimum acceptable standard (the outliers) had significantly  
17  
18 160 poorer performance on the BSIS real-life screening measures. PERFORMS outliers are readers whose  
19  
20 161 test performance falls more than one and a half times the inter-quartile range below the 25th  
21  
22 162 percentile in terms of either cancer detection rate in the PERFORMS test set or the area under the  
23  
24 163 curve of the receiver operating characteristic (ROC) analysis of their test set performance (or both).  
25  
26 164 For the purposes of this study, any reader who had been an outlier on any of the PERFORMS test-  
27  
28 165 sets included in three-year period, were allocated into an 'Outliers' group. The real-life cancer  
29  
30 166 detection rates, recall rates, and PPVs of PERFORMS outliers were then compared against those of  
31  
32 167 other readers using analysis of variance (ANOVA). The  $\alpha$ -level for statistical significance was set at  
33  
34 168 .05 for all analyses. Statistical calculations were performed using the IBM SPSS Statistics (version  
35  
36 169 23.0) statistical software (SPSS Inc., Chicago, IL).  
37  
38  
39  
40  
41  
42  
43  
44  
45

46 171 **Results**  
47  
48

49 172 *Participant Performance Overview*  
50  
51

52 173 In total, 452 participants (238 board certified radiologists, 193 radiographer readers, and 21 breast  
53  
54 174 clinicians) consented and were eligible to take part in the study. The mean cancer detection rate  
55  
56 175 from the BSIS real-life data was 7.79 per 1000 women screened (0.78%) with a mean recall rate of  
57  
58 176 5.29%. Each PERFORMS test set of 60 cases is heavily enriched with cancers; the number of cancer  
59  
60

1  
2  
3 177 cases varied between 34 and 38 for the PERFORMS sets included in this study. The mean cancer  
4  
5 178 detection rate in the PERFORMS test sets was 22.86% with a mean recall rate of 37.49%. A summary  
6  
7 179 of the BSIS real-life and PERFORMS performance measures for the participants is given in Table 1.  
8  
9

10  
11 180

### 14 181 *Test Measures Assessed from BSIS Real-life and PERFORMS Correlate*

15  
16  
17 182 BSIS real-life cancer detection rates, recall rates, and PPVs showed significant positive correlations  
18  
19 183 with the equivalent PERFORMS measures ( $n = 452$ ). Readers with a higher cancer detection rate in  
20  
21 184 real-life tended to have a higher cancer detection rate in PERFORMS (Pearson's Correlation:  $r =$   
22  
23 185  $0.179$ ,  $P < .001$ , two tails; Figure 2A). Readers with a higher recall rate in real-life screening tended  
24  
25 186 to have a higher recall rate in PERFORMS (Pearson's Correlation:  $r = 0.146$ ,  $P = .002$ , two tails; Figure  
26  
27 187 2B). PPV, the probability that a patient recalled following screening mammography has a confirmed  
28  
29 188 breast malignancy, reflects a combination of cancer detection rate and recall rate. Readers with a  
30  
31 189 higher PPV in real-life screening tended to have a higher PPV in PERFORMS (Pearson's Correlation:  $r$   
32  
33 190  $= 0.263$ ,  $P < .001$ , two tails; Figure 2C). It is noted that, as PPV is affected by the prevalence of the  
34  
35 191 disease, PPV in the test-set data was considerably higher than in the real-life data, reflecting the  
36  
37 192 difference in the prevalence of cancers in the two data-sets.  
38  
39

40  
41 193

### 44 194 *Comparison of Outliers and Nonoutliers*

45  
46 195 Outliers in the PERFORMS scheme were found to have significantly lower performance than other  
47  
48 196 readers in real-life screening in terms of cancer detection rate and PPV, but did not differ  
49  
50 197 significantly in terms of recall rate (Table 2). The mean BSIS real-life screening cancer detection rate  
51  
52 198 of PERFORMS outliers was 7.2 per 1000 women screened and was significantly lower than other  
53  
54 199 readers (non-outliers) where the cancer detection rate was 7.9 per 1000 women screened (ANOVA  
55  
56 200  $F(1, 450) = 9.78$ ,  $p = .002$ ,  $\omega = .014$ ) (Figure 3A). The mean BSIS real-life screening recall rate of  
57  
58 201 PERFORMS outliers was 5.5% and was not different from that of other readers who had a mean of  
59  
60



1  
2  
3 202 5.3% (ANOVA  $F(1, 450) = 0.67, P = .415, \omega = .003$ ) (Figure 3B). The mean BSIS real-life screening PPV  
4  
5 203 of PERFORMS outliers was 0.14% and was significantly lower than the non-outlier group who had a  
6  
7 204 mean PPV of 0.17% (ANOVA  $F(1, 450) = 7.75, P = .006, \omega = .012$ ) (Figure 3C).  
8  
9

10  
11 205

## 12 13 14 206 **Discussion**

15  
16  
17 207 This study was designed to determine if performance in the PERFORMS test set scheme reflected  
18  
19 208 BSIS real-life performance. Test set performance demonstrated significant positive correlations with  
20  
21 209 the BSIS real-life performance metrics produced by the UK screening programme, i.e. cancer  
22  
23 210 detection rate( $r = 0.179, P < .001$ ), recall rate( $r = 0.146, P = .002$ ), PPV( $r = 0.263, P < .001$ ) all showed  
24  
25 211 strong correlations. For breast cancer screening to be successful, cancer detection rates need to be  
26  
27 212 optimized, but at the same time recall rates need to be kept as low as possible to avoid false positive  
28  
29 213 interpretation and recalls. There will always be a trade-off between recalling women for further  
30  
31 214 investigation and detecting cancers, which is reflected in the PPV. Recall rates act as a proxy for  
32  
33 215 specificity in real-life screening, due to the difficulty in identifying true negatives and false negatives  
34  
35 216 at the time of reading. However, recall rates are not a perfect measure of specificity. Recall rates  
36  
37 217 need to be interpreted in conjunction with cancer detection – both low and high recall rates would  
38  
39 218 be acceptable in the context of high cancer detection, whereas in isolation extreme recall rates may  
40  
41 219 raise concerns about a reader's performance.  
42  
43  
44

45  
46 220 Correlation between BSIS real-life recall rates and PERFORMS correct recall rates was the least strong  
47  
48 221 of the performance metrics, although it did reach statistical significance ( $r = 0.146, P = .002$ ). One of  
49  
50 222 the criticisms of test sets is that reading behaviour may be altered. This weaker correlation is probably  
51  
52 223 not surprising as it has previously demonstrated that recall rates are particularly prone to this  
53  
54 224 'laboratory' effect, as readers know that flagging a patient for recall will have no impact on patient  
55  
56 225 care (4).  
57  
58  
59  
60

1  
2  
3 226 Previous studies comparing test-set and real-life performance have shown consistently positive  
4  
5 227 relationships, albeit weak in some instances (9-11). One of the strengths of this study is that it has  
6  
7 228 been possible to compare real-life performance data with results from a test-set scheme in a large  
8  
9 229 group of readers. Soh et al reported reasonable levels ( $P < .01$ ) of agreement between actual clinical  
10  
11 230 reporting and test set conditions, although increased sensitivity was seen under test set conditions  
12  
13  
14 231 (11). This study of 452 participants demonstrated much stronger associations than a previous smaller  
15  
16 232 study of 40 readers from one UK region taking part in the same PERFORMS scheme in 2005 and 2006  
17  
18 233 (10). PPV of recall demonstrated the strongest correlation between BSIS real-life and PERFORMS data  
19  
20 234 for all participants. PPV is one of the most useful measures of performance (12).

21  
22  
23  
24 235 Real-life performance data is often considered the reference standard. However, the accuracy of  
25  
26 236 sensitivity and specificity of real-life breast cancer screening data is problematic (13). Reader  
27  
28 237 sensitivity, which is defined as the proportion of patients with breast cancer reported as positive, is  
29  
30 238 not known for several years until interval cancer data becomes available and even then real life data  
31  
32 239 may not be updated to reflect this. Due to this unavoidable time lag, the opportunity to introduce  
33  
34 240 timely interventions to improve performance is lost. Similarly, when measuring specificity as the  
35  
36 241 proportion of disease-free patients reported as negative, a truly negative mammogram will not be  
37  
38 242 apparent until after the next screening round at the earliest. One of the advantages of test sets like  
39  
40 243 PERFORMS is that normal, benign, and malignant cases with known, biopsy proven outcomes and  
41  
42 244 appropriate follow up can be selected for inclusion, providing potentially more accurate performance  
43  
44 245 metrics. For instance, when choosing cases for PERFORMS, a normal case will only be included if the  
45  
46 246 mammogram at the next screening round three years later is also normal.

47  
48  
49  
50  
51 247 One of the key functions of measuring performance is to identify potential problems at the earliest  
52  
53 248 opportunity to allow interventions to change practice. Real life data is by its very nature retrospective.  
54  
55 249 Cancer detection rates of around 7-8 per 1000 women screened mean that an individual reader is  
56  
57 250 exposed to relatively few cancers each year. Consequently, it can be difficult to identify poor  
58  
59  
60

1  
2  
3 251 performance because of the statistical instability from the relatively small number of cancer cases,  
4  
5 252 similar problems are encountered when measuring performance in NHSBSP screening centres with  
6  
7 253 the smallest number of clients (14). BSIS audit data are combined over a three-year period to improve  
8  
9 254 the statistical robustness of the performance measures, but even so many years of poor performance  
10  
11 255 may occur before this becomes apparent through clinical audit, resulting in potential harm to the  
12  
13 256 screening population. For many years the PERFORMS scheme has flagged up poor performance  
14  
15 257 outliers where metrics have deviated significantly from the mean. Individuals and the regional quality  
16  
17 258 assurance office are notified so that corrective measures can be instigated such as reviewing practice  
18  
19 259 or further training. PERFORMS has the potential to identify under performance at a much earlier  
20  
21 260 stage than real-life data, perhaps even before a reader takes part in the screening programme as part  
22  
23 261 of an end of training or pre-employment assessment. If test sets are to be used in this way, then it is  
24  
25 262 crucial that the results are validated against real-life data. In this study, being a poor performance  
26  
27 263 outlier in PERFORMS was able to predict poor real-life performance with outliers have significantly  
28  
29 264 poorer real-life cancer detection rate and PPV of recall compared to the non-outlier group of readThis  
30  
31 265 study does have limitations. Nearly 20% of PERFORMS participants (124 readers) declined to have  
32  
33 266 their data used and so this has to be considered a potential source of bias. Further work is needed to  
34  
35 267 understand if this group had any particular characteristics.  
36  
37  
38  
39  
40  
41

42 268 In conclusion, there are significant correlations between real-life readers' performance in a breast  
43  
44 269 screening programme and their performance on metrics generated from a test-set based assessment  
45  
46 270 scheme such as PERFORMS. Readers' positive predictive value of recall in real-life screening and the  
47  
48 271 test-sets showed the strongest correlations. The use of test-set based assessment schemes has the  
49  
50 272 potential to predict and identify potential poor performance outliers in real-life screening, enabling  
51  
52 273 corrective measures to be implemented in a timely fashion.  
53  
54  
55

56 274

59 275 **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. Eggin 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](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.

- 1  
2  
3 7. Programme Specific Operating Model for Quality Assurance of breast screening Programmes.  
4  
5 Government UK website. [https://www.gov.uk/government/publications/breast-screening-](https://www.gov.uk/government/publications/breast-screening-programme-specific-operating-model)  
6  
7 [programme-specific-operating-model](https://www.gov.uk/government/publications/breast-screening-programme-specific-operating-model) Published July 2017 . Accessed August 15, 2019  
8  
9
- 10  
11 8. Quality assurance guidelines for breast cancer screening radiology. NHS Cancer Screening  
12  
13 Programmes website.  
14  
15 [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file](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)  
16  
17 [/764452/Quality\\_assurance\\_guidelines\\_for\\_breast\\_cancer\\_screening\\_radiology\\_updated\\_Dec\\_201](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)  
18  
19 [8.pdf](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  
20  
21
- 22  
23 9. Soh BP, Lee W, Kench PL, Reed WM, McEntee MF, Poulos A, et al. Assessing reader performance in  
24  
25 radiology, an imperfect science: Lessons from breast screening. *Clin Radiol* 2012; 67(7):623–28. doi:  
26  
27 10.1016/j.crad.2012.02.007  
28  
29
- 30  
31 10. Scott HJ, Evans A, Gale AG, Murphy A, Reed J. The relationship between real life breast screening  
32  
33 and an annual self-assessment scheme. Proceedings of SPIE, Medical Imaging 2009: Image Perception,  
34  
35 Observer Performance, and Technology Assessment; 72631E. doi: 10.1117/12.811003. Published  
36  
37 March 12, 2009. Accessed September 22, 2019  
38  
39
- 40  
41 11. Soh BP, Lee W, McEntee MF, Kench PL, Warren M., Heard R, et al. Screening Mammography: Test  
42  
43 Set data can reasonably describe actual clinical reporting. *Radiology* 2013; 268:46–53. doi:  
44  
45 10.1148/radiol.13122399  
46  
47
- 48  
49 12. Bennett RL, Blanks RG. Should a standard be defined for the Positive Predictive Value (PPV) of  
50  
51 recall in the UK NHS Breast Screening Programme? *Breast J* 2007; 16(1):55–9. doi:  
52  
53 10.1016/j.breast.2006.05.008  
54  
55
- 56  
57 13. Rutter CM, Taplin S. Assessing mammographers' accuracy: a comparison of clinical and test  
58  
59 performance. *J Clin Epidemiol* 2000; 53(5):443-450. doi: 10.1016/S0895-4356(99)00218-8  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

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

1  
2  
3 **Tables**  
4  
5

6 **Table 1: Summary of Real-life and PERFORMS Performance Measures**  
7

| Values                    | Real-life         |  |                    | PERFORMS             |                           |                      |
|---------------------------|-------------------|--|--------------------|----------------------|---------------------------|----------------------|
|                           | PPV (%)           | Cancers detected per 1000 women ( <i>n</i> ) | 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         |

8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32 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.  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

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

| Values                    | Real-life performance metrics                                   |                 |                 |                 |                               |                   |
|---------------------------|---|-----------------|-----------------|-----------------|-------------------------------|-------------------|
|                           | Number of cancers detected per 1000 women screened ( <i>n</i> ) |                 | Recall rate (%) |                 | Positive predictive value (%) |                   |
|                           | PERFORMS Outlier Status (2013-2016)                             |                 |                 |                 |                               |                   |
|                           | 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        |
| <i>P</i> 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.



1  
2  
3 **Figure 1:** Flowchart shows enrolment of readers into the study.  
4

5 **Figure 2.** Plots show correlation between **(a)** cancer detection rates, **(b)** recall rate, and **(c)** positive predictive  
6 value in real life and the PERFORMS tests sets.  
7

8 **Figure 2A:** Graph shows correlation between cancer detection in real life and in the PERFORMS test sets.  
9

10 **Figure 2B:** Graph shows correlation between recall rate in real life and in PERFORMS test sets.  
11

12 **Figure 2C:** Graph shows correlation between positive predictive value (PPV) in real life and PPV in the  
13 PERFORMS test sets.  
14

15  
16 **Figure 3:** A total of 396 non-outliers and 56 outliers were assessed for their cancer detection rates per 1000  
17 women, recall rates, and positive predictive value (PPV). Box-and-whisker plots show **(a)** real-life cancer  
18 detection rates, **(b)** real-life recall rates, and **(c)** real-life PPVs based on whether or not readers were an “outlier”  
19 in the PERFORMS test sets. The 95% confidence limits are shown on each plot.  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60