BJR

https://doi.org/10.1259/bjr.20190908

Received: 25 October 2019 | Revised: 22 May 2020 | Accepted: 29 May 2020

Cite this article as:
Gale A, Chen Y. A review of the PERFORMS scheme in breast screening. *Br J Radiol* 2020; **93**: 20190908.

REVIEW ARTICLE

A review of the PERFORMS scheme in breast screening

[1]ALASTAIR GALE, PhD FBPS CPsychol FCIEHF HonFRCR and [2]YAN CHEN, PhD HonMRCR

[1]Department of Computer Science, Loughborough University, Loughborough, UK
[2]Division of Cancer and Stem Cells, University of Nottingham, Nottingham, UK

Address correspondence to: Professor Alastair Gale
E-mail: *A.G.Gale@lboro.ac.uk*

ABSTRACT:

This review details the aetiology of the PERFORMS self-assessment scheme in breast screening, together with its subsequent development, current implementation and future function. The purpose of the scheme is examined and the importance of its continuing role in a changing screening service described, together with current evolution.

INTRODUCTION

Every year more than 2 million females have breast cancer screening in the UK,[1] while approximately one in eight are diagnosed with breast cancer during their lifetime.[2] Breast screening was introduced in the UK in 1988 following the Forrest report[3] which recommended screening every 3 years for all females aged 50 to 64. Screening is crucial to early breast cancer detection; however, the low incidence rate during screening compounds any difficulty readers have in identifying early signs of abnormality readily and also for them to be aware of details of their own performance. Females aged 50 to 70 are currently invited for screening[4,5] with further age extensions of 47 to 73 years under investigation via the AgeX trial.[6]

Screening is a two-stage process. A female is first imaged using two view full-field digital mammography (FFDM), and her case reported as either "return to screen" or if the mammographic appearance is suspicious then she is "recommended for assessment". Subsequent further imaging and investigations at assessment determine whether her mammographic appearance is abnormal and what action to take.

Since its inception, the UK screening programme has always aimed for a low recall for assessment rate which has been found to be circa 4.56%[7] and to maintain this level necessitates that high-quality abnormality detection decisions are made concerning which cases merit recall. In 2016 in England, some 7.8% of prevalent round females were recalled for assessment with 3% of incident screened females recalled.[5] A breast screening reader, typically either a radiologist or specially trained advanced practitioner radiographer, gains rapid feedback on whether their decisions to recommend a case for assessment were appropriate. However, with a 3-year screening interval, there is no swift feedback for a case considered normal or definitely benign and consequently not recalled for assessment. In such instances, appropriate feedback to the reader cannot occur until the female presents for the subsequent screening round 3 years later or if she presents with an interval cancer in the intervening period.

To help maintain and improve radiological skills in determining which cases truly require recalling the PERFORMS scheme[8,9] was developed. In the UK, each breast screening reader has been recommended to report 5000 cases per year,[10] thereby on average is only presented with some 35 cancer cases annually. By also examining the PERsonal perFORmance in Mammographic Screening (PERFORMS) test sets, they could additionally read 80–90 challenging cancer cases per annum.

THE PERFORMS SCHEME

The PERFORMS scheme was invented and developed by A.G. Gale, working with E.J. Roebuck[11] one of the pioneers of screening in the UK. It began as a small-scale unfunded research project in the East Midlands region of England with the motivation of modelling variations in radiologists' skills in identifying key mammographic appearances of early breast cancer. This led to its development as a self-assessment scheme which was then gradually established nationally in parallel with the introduction of UK nationwide breast cancer screening. This development was in conjunction with the Royal College of Radiologists (RCR) and the UK NHS Breast Screening Programme which originally funded the programme nationally. The scheme has been implemented across the UK continuously for over 30

years. It is the leading and, first, national self-assessment scheme in radiology.[12–14] It provides an individual with self-assessment on imaging decisions as judged against: expert radiological opinion, known case pathology, peer opinions, and also provides personalised training and additional screening programme quality assurance functions.

Participation in the scheme has always been highly recommended by the RCR[15] and more recently has become mandated by Public Health England (PHE)[16] with readers receiving appropriate professional CME (Continuous Medical Education) points for taking part. It has also been independently recommended to improve reader skills following an external review of a screening incident which had resulted in some 61 cancers being missed.[17]

Screening in the UK originally used a single Medio-Lateral Oblique (MLO) mammographic film of each breast and so the PERFORMS scheme was first designed and deployed using radiographical copies of single MLO images, together with a paper-based response system where readers recorded their decisions on the cases.[14] This information was mailed back for analyses, with subsequent detailed feedback again being returned to the reader. Inevitably this offered readers delayed and simplistic performance analyses. Sets of screening cases were couriered around the UK, coupled with reporting books which readers completed. Two view screening (MLO and CC; Cranio-Caudal) was introduced in 1995 and the scheme developed in parallel. It subsequently evolved using various personal computer systems to provide rapid feedback to readers immediately after they had read a set of test mammographic cases. As all screening centres progressed after 2010 in using two-view FFDM, then a cloud-based approach for the scheme became feasible.

CURRENT IMPLEMENTATION

De-identified recent potential test FFDM cases are sourced widely from UK screening centres, complete with the original radiological opinion and any pathological report. A panel of over 10 expert breast radiologists then individually identify key mammographic features in each case and rate case density, case difficulty and other factors, determining whether each case is challenging enough (*i.e.,* approximately 75% of participants could possibly report it incorrectly) to be included in the scheme. In collaboration with further experts, we then derive a consensus opinion per case. Finally, set of cases are constructed which include a mixture of normal, benign and malignant cases with various features including subtle or interval cancers, bilateral appearances, multifocal tumours, etc. Case sets are updated annually.

PERFORMS is now[18] a complex online training scheme with extensive educational feedback and is undertaken annually by all (almost 1000) screening readers nationally as well as symptomatic radiologists (these are radiologists who do not participate in the national screening programme but do read symptomatic breast cases). The scheme has been developed to be as simple and transparent in use to readers as possible. With a range of different vendor workstations in use across the UK, it is imperative that readers employ their own workstations and associated familiar viewing software to examine the FFDM test cases, so that the recorded readers' performances reflect their everyday screening behaviour. Readers simply use whatever workstation software they usually employ to view and examine the test cases and then use the PERFORMS App, typically running on a laptop or other computing device beside the workstation (although it can also run on the workstation itself if this is internet enabled), to report the cases and receive detailed feedback. The scheme then easily accommodates any vendor software updates or change in vendor supplier by a breast screening centre. It has purposefully not sought to be a standalone DICOM (Digital Imaging and COmunications in Medicine) viewer and reporting application running on the reader's workstation.

Every reader has their own PERFORMS confidential web portal which provides detailed ongoing performance data and from where they can log in to the dedicated web-based reporting App; download numerous sets of test cases; report these cases; and receive feedback. Using the App, readers locate and identify several key mammographic appearances, rating each feature in terms of level of suspicion. They also classify whether to recall or not recall each breast image using the UK RCR screening classification rating scale[19] or when the scheme is being used outside the UK, they use the appropriate Breast Imaging Reporting and Data System (BIRADS) rating scale.[20] The two scales having been shown to be well related.[21]

The App provides immediate feedback and readers can revisit any, or all, cases and examine annotated expert radiological pictorial and textual feedback as well as associated relevant pathology information. After taking part, they can, via their web portal, view personalised performance reports; enter into online peer and expert discussions on cases, and other associated activities. Once all readers have completed a scheme then all data are statistically analysed, and each reader receives detailed and anonymised web-based reports regarding their performance in comparison with their colleagues nationally. Management reports for the screening programmes are also regularly produced.

Of necessity, the large number of readers need to take part over several months as they schedule participation around their normal screening workload. Consequently, daily updated comparative and anonymised peer review is available where a user can keep revisiting the website to compare their own case decisions to an ever-growing number of decisions of their peers. Various interactive performance tables are available which highlight cases which were most misreported (*e.g.,* false negatives, false positives) as well as other factors such as most missed abnormalities. Ongoing current work is examining the possible use of augmented reality as part of enhanced educational feedback to readers.[22]

A key aspect of the PERFORMS scheme is that it can statistically identify mild and severe underperforming outliers and then follow-up these individuals with suggestions of how to improve their performances (which can entail detailed visual search recording in our laboratory) as well as monitor them on

future test sets. Typically, a few outliers are found annually out of the hundreds of readers taking part. In general, substandard performance is found to relate to an individual's clinical workload at the time of reading the case set. As all reader responses are timed then any outlying performance can be investigated to firstly determine how long the user spent reading whole case sets, reading specific cases and the time of day when reporting. Any outlier typically is found to improve their performance on subsequent iterations of the scheme. Previously, outliers have only been able to be identified when all readers have taken part in a particular PERFORMS case set. However, we have successfully demonstrated that we can now statistically predict potential outliers only after a few readers have taken part which means that potential outliers can be identified earlier whilst a scheme is being implemented and so given help much sooner.[23]

The PERFORMS scheme is deliberately designed to allow readers both to participate when they like and also to start, stop and restart at will. This allows readers to fit taking part into their clinical work and at a time that suits them. However, extending the test set reading over time may also introduce other factors into the performance measurements. In practice, the majority of UK participants read the 60 test cases in one continuous reporting session. We have examined data on readers taking breaks while taking part and found little or no effect whether they took mini-breaks when reading a case set or whether they spread reading a case set over clinical reporting sessions. When readers read the case set in one complete session, without mini-breaks, more errors occurred as reading time increased.[24] This indicates possible effects of fatigue.

PERFORMS and real-life screening data

An important issue is how an individual's performance on the PERFORMS scheme relates to their performance in screening practice. Inherently, reporting selected test cases is essentially the same task as routine everyday reporting and so similar performance values between the two situations might be expected. However, differences between the two situations exist that must be acknowledged when interpreting information such as the reader's knowledge before reporting that these are carefully selected test cases. While the actual variables recorded in PERFORMS are based upon the routine measures reported in everyday UK breast screening, the particular exemplar cases used in the scheme are carefully selected to be challenging enough to be able to tease out potential performance differences between participants. Each set of difficult test cases is carefully derived by a panel of experts as case difficulty can be very subjective.[25] Apart from the test cases being more difficult than found in routine screening, the case sets are necessarily weighted with more benign and malignant cases so as to elicit variance in performance measures. In the UK, with an individual reading 5000 cases a year, we estimate that in one year our scheme would give them the equivalent experience of challenging abnormal cases as they would experience when reading several years' worth of routine screening cases.

In the UK, readers' usual screening performance is monitored and audited annually, with data concerning every screening centre published. This provides important information for individuals concerning their performance. However, using the PERFORMS test sets additionally to assess aspects of their performance objectively is important, not only because these data can give rapid feedback on an individual's skills but also because it is virtually impossible to reliably assess an individual's performance based solely on annual real-life data. Instead, their real-life screening performance must be based on several years of screening data in order to encompass them reporting a representative number of cancer cases. This is because the number of abnormal cases which a reader will see per annum is approximately 35 per 5000 cases screened (based on an approximate cancer detection rate of 7 per 1000 screened cases). This means that an individual's annual sensitivity measure would be based solely on a very small number of abnormal cases out of the 5000 cases examined. Therefore, picking up, or missing, a very small number of these abnormal cases would lead to a relatively large difference in the sensitivity measure in their annual performance assessment.

Over the years of implementation various comparisons have been undertaken between the data from the PERFORMS scheme and real-life screening. This has included comparisons between the mammographic features missed in PERFORMS and those that were recorded in an interval cancer database. The finding that similar proportions of features were missed in both situations demonstrated that readers had comparable difficulty with these features, both in real life and in the scheme. Additionally, readers' real-life screening recall rates correlated with PERFORMS measures of correct recall and correct return to screen.[26] Subsequently the relationship between different types of readers in one UK health region was studied using data from the national NHS National Breast Screening System (NBSS) database and comparable PERFORMS data for the similar time frame. Significant positive correlations were found between the PERFORMS data and most of the comparable real-life measures.[27] More recently, the Breast Screening Information System (BSIS) has been introduced across England, which provides screening programme information as well as individuals' real-life performance data over 3-year periods. Consequently, real-life screening performance data from over 450 screeners in England have been examined[28] over such a 3-year period and compared to their PERFORMS data over a similar time frame. Significant positive correlations were found between real-life cancer detection rates, recall rates and PPVs (Positive Predictive Value) and the equivalent PERFORMS measures.

Taken together, all these data demonstrate that an individual's performance on the PERFORMS scheme can be seen as a useful surrogate indicator of their real-life screening performance. Importantly, this measure can be achieved quickly, without having to wait years for real-life data to be amassed, and so if an underperforming individual is found on the PERFORMS scheme then additional training can readily be offered to improve their real-life screening performance.

Factors which underlie screening performance

The theoretical underpinning of the performance measures used in the PERFORMS scheme is based upon an active vision approach to perception.[29] Although mistakes occur in any

task for a variety of reasons,[30] the key to understand radiological performance is basically how individuals detect and identify known features indicative of normal, benign or abnormal appearances.[31,32] Visual search of the medical image is a key factor in this visuo-cognitive process and is complex,[33] primarily being composed of a series of saccadic eye movements which serve to shift the fovea to fixate upon specific areas in the image. As saccades are ballistic then these movements are pre-planned. The schematic map's[34] theoretical approach well describes this process where the first fixation on the image establishes an initial image gist[34] which is followed by a series of eye fixations driven by the developing cognitive schema of the image at each subsequent eye fixation which determines the next fixation location and so on.[35,36] Researchers[37–39] have argued that medical image inspection involves both global and focal processes underlying such medical image search behaviour. Many studies investigating visual search and expertise in different scenarios[40] have shown that experienced individuals fixate important image areas faster and for longer periods than naïve observers, who make more and shorter disorganised fixations per unit time.[41] Such an approach describes both the acquisition of radiological knowledge as well as how experienced radiologists' image examination is subsequently almost automatically and quickly performed.

In breast screening research, most interest is in false-negative decisions where an abnormality is missed. Such errors can be classed as due to visual search, detection or classification.[37] Various research studies have demonstrated that, depending on the study, approximately 24–30% of false-negative errors are due to errors of visual search, some 25% perceptual and 45–52% cognitive.[37,3838] The PERFORMS scheme enables clarification of any errors as due to detection or classification. By inspecting a reader's feature identifications against the expert panel decisions, it is possible to assess whether they have identified appropriate features or not (detection error) and if appropriately identified have they then reported them suitably (interpretation error). In order to determine whether any errors are due to visual search *per se,* then the eye movements of the reader need to be recorded and this is something we have done with outliers to give them better insight into this aspect of their performance.

The type of reporting error can be directly related to the type of mammographic features[42–45] identified,with some features being more likely to be undetected and others misinterpreted. In addition to the importance of key mammographic features, breast density[46] is an important underlying factor which increases the risk of breast cancer and can also mask its appearance. High breast density is associated with a greater chance of developing breast cancer, but unfortunately dense images are much harder to interpret, therefore making it more difficult to identify breast cancer accurately.[47] We have examined how density was reliably reported across individuals and across different participant groups[48] with better agreement being found amongst radiologists on case density ratings as compared to advanced practitioners.

The usual approach to determining correct feature identification is to assume a circular area of interest (AOI) centred around the abnormality, even when the abnormality has been clearly demarcated by experts as irregularly shaped. This is a reasonable working approach and a response within the AOI is then taken as a correct abnormality identification and detections outside the AOI marked as false positives. However, an improved technique has been developed[23] which more closely follows the contours of an abnormality and therefore can more accurately determine whether a reader's response should truly be considered as a correct detection or not.

Expertise and reading case volume

The development of expertise in breast screening is a complex process, comprising both reading a large number of cases per annum and also years of reporting cases. Several studies[48–51] have tried to tease apart whether it is years of experience or volume of cases which is the more important by examining data from the scheme. In the UK, all readers are recommended to read at least 5000 cases per annum as a means of developing and maintaining expertise in the domain.[10] By doing so they then inspect many normal cases and develop an appreciation of the wide range of normal appearances. Increasing expertise serves to develop appropriate cognitive schema for normal and abnormal appearances which then reduces the time required to examine a case as experienced readers search and concentrate upon high probability areas of an image where an abnormality may reside, as well as develop experience in identifying early features of abnormality.

A very different approach is in the USA where the continuing experience requirement is to read 960 cases over 2 years.[52] The role of case volume has been examined by studying UK and American radiologists reading PERFORMS cases, with the American group split into high-, medium- and low-volume readers. Data demonstrated that the lower volume readers had significantly lower sensitivity than high-volume readers or the UK radiologists supporting the need for high volume.[53] Further comparative studies have shown largely similar performances between radiologists from the two countries, as well as from Europe, but with differences in recall decisions interpreted as reflecting different health management approaches.[54–56]

Emergent future roles

PERFORMS was developed with the aim of helping readers identify early signs of breast cancer. As the UK screening programme has matured such a need has been maintained as more new readers, particularly advanced practitioners, join the screening programme and PERFORMS data have shown how well such practitioners can perform. Not surprisingly, mean values for cancer detection rate and correct recall rate improve as individuals' experience in screening grows. Readers learn many abnormality identification skills fairly quickly in their first year of screening.[57]

High cancer detection performance should be equivalent to good correct recall performance; however, this is not always the case as good abnormality detection can be achieved by simply over reading cases which simply increases assessment numbers. While PERFORMS has successfully aided readers to identify early difficult abnormality appearances, it is important in actual

screening that the number of cases recommended for assessment are both minimised and fully appropriate. Consequently, PERFORMS is developing targeted test sets of cases which emphasise key benign appearances with the aim of increasing knowledge of the range of benign presentations which do not actually merit recall.

The PERFORMS scheme initially grew out of earlier work[58] which had developed a computerised decision aid based upon accurate radiological identification of a small number of key mammographic features. However, such accurate feature identification across readers is not feasible, due to intra- and inter-individual variability in reporting, which limited the generalisability of such an approach. More robust computer-aided detection and decision systems have been developed with variable success, tending to target-specific mammographic features.[59] A UK trial (CADET II) demonstrated that CAD systems could potentially replace a reader in a double reading scenario.[60] However, CAD approaches have never had any great impact or widespread support within the UK screening community.

In marked contrast, the subsequent growth of artificial intelligence (AI) and machine learning approaches has come at a time when the UK is struggling to recruit radiologists, and these may therefore potentially offer considerable help to the screening programme.[61] There are several ways in which an appropriate AI approach could be implemented in UK breast screening, the most promising may be to filter out a large number of normal cases so reducing the readers'workload.[62]

AI approaches are developed using a large database of training cases and then the developed algorithms are tested using a separate large set of test cases with the algorithm's performance generally being taken as its success on these test cases. However, there is a growing realisation of the need for a third stage of AI algorithm validation which is where the PERFORMS scheme is valuable. By applying AI algorithms to sets of cases where the clinical outcome is both known and which large numbers of radiologists have read, then it becomes possible to rate algorithms' performance against health professionals' skills.

Consequently, the PERFORMS AI Grand Challenge[63] has been established to allow developers to assess their software on the same challenging cases, both against other AI systems, and also against a large cohort of experienced breast screeners. AI developers can therefore directly see how their system would potentially perform in the real screening environment and possible purchasers can know whether systems are truly ready for market and fit for purpose.

The PERFORMS scheme was established specifically for mammography in breast screening. However, the approach has since been extended to encompass digital breast tomography (DBT). A current major UK trial (PROSPECTS) of DBT in screening is underway in which the scheme is involved. It is also embedded as part of a major EU H2020 international trial (MyPEBS) of personalized breast screening across Europe where the role of the scheme is to monitor performance across hundreds of readers from different countries over the course of the trial. Based on some aspects of the PERFORMS model, some other countries (*e.g.,* Australia, Netherlands) have subsequently developed their own breast screening monitoring approaches. The PERFORMS scheme is available internationally and has also been developed for other radiological domains (www.iperforms.com).

CONCLUSIONS

PERFORMS is a UK national self-assessment and educational scheme which has been implemented alongside the breast screening programme, providing peer review and quality assurance functions. The scheme quickly identifies individuals who are underperforming and facilitates their improvement. Through a process of continuous development, the scheme maintains advantage to readers and the screening programme. The scheme is implemented internationally as well as being developed across other radiological domains.

ACKNOWLEDGMENT

This work has been supported by the NHS Breast Screening Programme; Public Health England; the Scottish Breast Screening Programme; Breast Test Wales, and the Public Health Agency, Northern Ireland. PERFORMS is a registered trademark.

REFERENCES

1. https://www.cancerresearchuk.org/about-cancer/breast-cancer/screening/breast-screeningaccessed 4 Sept 2019
2. http://digital.nhs.uk/Breast Screening Programme Englandaccessed 4 Sept 2019
3. Forrest APM. *Report to the Health Ministers of England, Wales, Scotland and Northern Ireland*. London, UK: HMSO; 1986.
4. https://www.nhs.uk/NHSEngland/NSF/Documents/Cancer%20Reform%20Strategy.pdfaccessed 4 Sept 2019
5. NHS2017https://digital.nhs.uk/catalogue/PUB23376accessed 4 Sept 2019
6. Moser K, Sellars S, Wheaton M, Cooke J, Duncan A, Maxwell A, et al. Extending the age range for breast screening in England: pilot study to assess the feasibility and acceptability of randomization. *J Med Screen* 2011; **18**: 96–102. doi: https://doi.org/10.1258/jms.2011.011065
7. Burnside ES, Vulkan D, Blanks RG, Duffy SW. Association between screening mammography recall rate and interval cancers in the UK breast cancer service screening program: a cohort study. *Radiology* 2018; **288**: 47–54. Epub 2018 Apr 3. doi: https://doi.org/10.1148/radiol.2018171539
8. Gale AG, Walker GE. Design for performance: quality assessment in a national breast screening programme. In: Lovesay E, ed. *Ergonomics - design for performance* . . London, UK: Taylor & Francis; 1991. pp. 197–2021991.
9. Gale AG. Performs – a self-assessment scheme for radiologists in breast screening. *Seminars in Breast Disease: Improving and monitoring mammographic interpretative skills* 2003; **6**: 148–52.

10. Royal College of Radiologists. *Quality Assurance Guidelines for Radiologists*. Royal College of Radiologists, UK; 1990.
11. Roebuck EJ. Mammography and screening for breast cancer. *Br Med J* 1986; **292**: 223–6. doi: https://doi.org/10.1136/bmj.292.6515.223-a
12. CowleyH. Gale AG: Minimising human error in the detection of breast cancer. In: Robertson S. A, ed. *Contemporary Ergonomics 1996*. London, UK: Taylor and Francis; 1996. pp. 379–84.
13. Gale AG. Scott H. Measuring Radiology Performance in Breast Screening. In: Michell M, ed. *Contemporary Issues in Cancer Imaging – Breast Cancer*. Cambridge: UP, Cambridge; 2010. pp. 29–45.
14. Gale AG. Maintaining quality in the UK breast screening program. In: Manning D. J, Abbey C, eds. *SPIE Medical Imaging 2010: Image Perception, Observer Performance, and Technology Assessment*; 2010. pp.. . 762702-1–762702-11.
15. Royal College of Radiologists. *Guidance on screening and symptomatic breast imaging*. Third edition. London: The Royal College of Radiologists; 2013.
16. NHS public health functions agreement 2018-19 Service specification no. *24*. London: Breast Screening Programme; 2018.
17. Burns FG. An independent external review of the breast screening unit at East Lancashire NHS trust. *Burnley, UK. East Lancashire NHS Trust* 2011;.
18. Gale AG. Chen Y. PERFORMS – Performance assessment using standardised data sets. In: Samei E, Krupinski E, eds. *The Handbook of Medical Image Perception and Techniques*. second edition: Cambridge University press; 2018.
19. Maxwell AJ, Ridley NT, Rubin G, Wallis MG, Gilbert FJ, Michell MJ, et al. The Royal College of radiologists breast group breast imaging classification. *Clin Radiol* 2009; **64**: 624–7. doi: https://doi.org/10.1016/j.crad.2009.01.010
20. Sickles EA, D'Orsi CJ, Bassett LW.et al. ACR BI-RADS® Mammography. In: *ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System*. Reston, VA: American College of Radiology; 2013.
21. Taylor K, Britton P, O'Keeffe S, Wallis MG. Quantification of the UK 5-point breast imaging classification and mapping to BI-RADS to facilitate comparison with international literature. *Br J Radiol* 2011; **84**: 1005–10. doi: https://doi.org/10.1259/bjr/48490964
22. Tang Q, Dong L, Chen Y, Gale AG. The implementation of an AR (Augmented Reality) approach to support mammographic interpretation training – an initial feasibility study. In: Kupinski M. A, Nishikawa R. M, eds. *SPIE Medical Imaging 2017: Image Perception, Observer Performance, and Technology Assessment*; 2017.
23. Dong L, Chen Y, Gale AG, Chakraborty DP. A potential method to identify poor breast screening performance? In: Abbey C. K, Mello-Thoms C. R, eds. *SPIE Medical Imaging 2012: Image Perception, Observer Performance, and Technology Assessment*; 2012. pp.. ; 831819-1–831819-8..
24. Cowley HC. Gale A.G.: Time of Day Effects on Mammographic Film Reading Performance. In: Kundel H, ed. *Medical Imaging 1997: Image Perception. SPIE*; 1997. pp. 212–21.
25. Scott HJ, Gale AG. . Breast screening: when is a difficult case truly difficult and for whom? In: Eckstein M. P, Jiang Y, eds. *Image Perception, Observer Performance, and Technology Assessment*: Proceedings of SPIE; 2005. pp.. , 557–65.
26. Cowley HC, Gale AG. Breast cancer screening: comparison of radiologists' performance in a self-assessment scheme and in actual breast screening. In: Krupinski E. A, ed. *Medical Imaging 1999: Image and Performance*. **3663**: Proceedings of SPIE; 1999. pp. 157–68.
27. Scott HJ, Evans A, Gale AG, Murphy A, Reed J. The relationship between real life breast screening and an annual self-assessment scheme. In: Sahiner B, Manning D. J, eds. *SPIE Medical Imaging 2009: Image Perception, Observer Performance, and Technology Assessment*. **7263**; 2009. pp.. , E1–9.
28. Chen Y, James JJ, Cornford EJ, Jenkins J. Relationship between mammography readers real-life performance and performance in a Test-set based assessment scheme in a national breast screening programme radiology. *Imaging Cancer*.
29. Findlay JM, Gilchrist ID. Active vision: the psychology of looking and seeing. *Oxford, UK. Oxford University Press* 2003;.
30. Reason J. *Human error*. Cambridge, UK.: Cambridge University Press; 1991.
31. Garland LH. On the scientific evaluation of diagnostic procedures. *Radiology* 1949; **52**: 309–28. doi: https://doi.org/10.1148/52.3.309
32. Yerushalmy J. The statistical assessment of the variability in observer perception and description of roentgenographic pulmonary shadows. *Radiol Clin North Am* 1969; **7**: 381–90.
33. Brogan D, Carr K, Gale A. G. eds. *Visual Search II*. London: Taylor and Francis; 1993 .
34. Mugglestone MD, Gale AG, Cowley HC, Wilson ARM. Diagnostic performance on briefly presented mammographic images in Kundel HL editor. image perception. *Proc. SPIE* 1995; **2436**: 106–16Republic of Belarus, 220030.
35. Peterson MA, Gillam B. *editors: In the mind's eye: Julian Hochberg on the perception of pictures, films, and the world*. Sedgwick HA: Oxford University Press; 2007.
36. Neisser U. Cognition and reality. *WH Freeman* 1976;.
37. Kundel HL, Nodine CF, Carmody D. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Invest Radiol* 1978; **13**: 175–81. doi: https://doi.org/10.1097/00004424-197805000-00001
38. Krupinski EA. Visual scanning patterns of radiologists searching mammograms. *Acad Radiol* 1996; **3**: 137–44. doi: https://doi.org/10.1016/S1076-6332(05)80381-2
39. Gale AG. Human response to visual stimuli. In: Hendee W, Wells P, eds. *Perception of Visual Information - second edition*. New York, USA: Springer Verlag; 1997. pp. 127–47.
40. Gale AG, Johnson F.editors *Theoretical and Applied aspects of Eye Movement Research : Advances in Psychology*. Amsterdam, Netherlands: Elsevier Science B.V. North-Holland; 1984.
41. Mallett S, Phillips P, Fanshawe TR, Helbren E, Boone D, Gale A, et al. Tracking eye gaze during interpretation of endoluminal three-dimensional CT colonography: visual perception of experienced and inexperienced readers. *Radiology* 2014; **273**: 783–92. doi: https://doi.org/10.1148/radiol.14132896
42. Savage CJ, Gale AG, Pawley EF, Wilson ARM. To err is human, to compute divine. In: Gale A. G, Astley S. M, Dance D. R, Cairns A. Y, eds. *Digital Mammography. Proceedings of the 2nd International Workshop on Digital Mammography*. Elsevier, Amsterdam; 1994. pp. 405–14.
43. Whatmough P, Gale AG, Wilson ARM. Do radiologists agree on the importance of mammographic features? In: Doi K, Giger M. L, Nishikawa R. M, Scmidt R. A, eds. *Digital Mammography '96. Elsevier*. Amsterdam; 1996. pp. 111–6.
44. Scott HJ, Gale AG, Hill S. How are false negative cases perceived by mammographers? Which abnormalities are misinterpreted and which go undetected? In: Manning D, Sahiner B, eds. *Image Perception, Observer Performance, and Technology Assessment*: Proceedings of SPIE ; 2008. pp.. 1–116917.13,.
45. Dong L, Chen Y, Gale AG. Breast screening: understanding case difficulty and the nature of errors. In: Abbey C. K, Mello-Thoms C. R, eds. *Proceedings of SPIE, vol 8673, Medical Imaging 2013: Image Perception, Observer*

Performance, and Technology Assessment; 2013. pp. 867316-1–8.
46. Boyd NF, Rommens JM, Vogt K, Lee V, Hopper JL, Yaffe MJ, et al. Mammographic breast density as an intermediate phenotype for breast cancer. *Lancet Oncol* 2005; **6**: 798–808. doi: https://doi.org/10.1016/S1470-2045(05)70390-9
47. Darker IT, Chen Y, Gale AG. Health professionals' agreement on density judgements and successful abnormality identification within the UK Breast Screening Programme. In: Manning D. J, Abbey C. K, eds. *Proceedings of SPIE, Vol. 7996, Medical Imaging 2011: Image Perception, Observer Performance, and Technology Assessment*; 2011. pp. 796604-1–796604-10.
48. Scott HJ, Gale AG, Wooding DS. Breast Screening Technologists: does real-life case volume affect performance? In: Chakraborty D. P, Eckstein M. P, eds. *Image Perception, Observer Performance, and Technology Assessment, Proceedings of SPIE* , ; 2004. pp. 399–406.
49. Scott HJ, Gale AG, Wooding DS. European Breast Screening Performance: does case volume matter? In: *Image Perception, Observer Performance, and Technology Assessment, D.P. Chakraborty & M.P. Eckstein (eds.) Proceedings of SPIE* , ; 2004. pp. 383–90.
50. Scott HJ. Gale a. *G.: Breast screening: PERFORMS identifies key mammographic training needs. B J Radiol* 2006; **79**: S127–33.
51. Scott HJ. Gale A.G.: How much is enough: factors affecting the optimal interpretation of breast screening mammograms. In: Jiang Y, Sahiner B, eds. *Image Perception, Observer Performance, and Technology Assessment*. **6515**: Proceedings of SPIE; 2007. pp.. ; 65150F-1–65150F-10.
52. American College of Radiology. 2017. Available from: http://www.acraccreditation.org/~/media/ACRAccreditation/Documents/Mammography/Requirements.pdf?la=en [Last accessed 4 Sept, 2019].
53. Esserman L, Cowley H, Eberle C, Kirkpatrick A, Chang S, Berbaum K, et al. Improving the accuracy of mammography: volume and outcome relationships. *J Natl Cancer Inst* 2002; **94**: 369–75. doi: https://doi.org/10.1093/jnci/94.5.369
54. Chen Y, Gale AG, Evanoff M. Performance differences across the Atlantic when UK and USA radiologists read the same set of test screening cases. In: Abbey C. K, Mello-Thoms C. R, eds. *Proc. SPIE Vol 8318, Medical Imaging 2012: Image Perception, Observer Performance, and Technology Assessment*. **831811**. pp., ; 831811-1–831811-7.
55. Chen Y, Gale AG, Evanoff M. Does routine breast screening practice over-ride display quality in reporting enriched test sets? In: Abbey C. K, Mello-Thoms C. R, eds. *SPIE Medical Imaging 2013: Image Perception, Observer Performance, and Technology Assessment*. **8673**; 2013.
56. Chen Y, Dong L, Nevisi H, Gale AG. The international use of PERFORMS mammographic test sets. In: Tingberg A, Lång K, Timberg P, eds.*Breast Imaging: 13th International Workshop, IWDM 2016, Lecture Notes in Computer Science, Springer*; 2016. pp. 130–5.
57. Nevisi H, Dong L, Chen Y, Gale AG. How quickly do breast screeners learn their skills? In: Kupinski M. A, Nishikawa R. M, eds. *SPIE Medical Imaging 2017: Image Perception, Observer Performance, and Technology Assessment*; 2017. https://doi.org/.
58. Gale AG, Roebuck EJ, Riley P, Worthington BS. Computer AIDS to mammography diagnosis. *B J Radiol* 1987; **60**: 887–91.
59. Khoo LAL, Taylor P, Given-Wilson RM. Computer-Aided detection in the United Kingdom National breast screening programme: prospective study. *Radiology* 2005; **237**); : 444–9Vol.. doi: https://doi.org/10.1148/radiol.2372041362
60. Gilbert FJ, Astley SM, Gillan MGC, Agbaje OF, Wallis MG, James JJ, et al. For the CADET II group) single reading with computer-aided detection for screening mammography. *Engl J Med* 2008; ; **359**: 1675–84October 16N2008.
61. Harvey H, Edith Karpati E. Khara G &Korkinof D, Ng A & Austin C &Rijken T, Kecskemethy P. *The Role of Deep Learning in Breast Screening Current Breast Cancer Reports* 2019; **11**: 17–22.
62. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Broeders M, Gennaro G, Clauser P, et al. Stand-Alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J Natl Cancer Inst* 2019; **111**: 916–22. doi: https://doi.org/10.1093/jnci/djy222
63. Chen Y. PERFORMS Grand Challenge.. Available from: https://performs.grand-challenge.org.