Received:
6 January 2014

Revised:
1 May 2014

Accepted:
7 May 2014

doi: 10.1259/bjr.20140029

Cite this article as:

Lewis SJ, Mello-Thoms CR, Brennan PC, Lee W, Tan A, McEntee MF, et al. Assessing the effect of a true-positive recall case in screening mammography: does perceptual priming alter radiologists' performance? *Br J Radiol* 2014;87:20140029.

FULL PAPER

Assessing the effect of a true-positive recall case in screening mammography: does perceptual priming alter radiologists' performance?

¹S J LEWIS, PhD, ¹C R MELLO-THOMS, PhD, ¹P C BRENNAN, PhD, ¹W LEE, MD, ¹A TAN, BAppSc (Hons), ¹M F MCENTEE, PhD, ²M EVANOFF, PhD, ¹M PIETRZYK, PhD and ¹W M REED, PhD

¹Medical Imaging Optimisation and Perception Group (MIOPeG), Discipline of Medical Radiation Sciences, Lidcombe, Faculty of Health Science, The University of Sydney, NSW, Australia

²American Board of Radiology, Tucson, AZ, USA

Address correspondence to: Dr Sarah Jayne Lewis

E-mail: sarah.lewis@sydney.edu.au

This paper was presented by Dr W Reed at the 2013 MIPS conference in Washington, DC.

Objective: To measure the effect of the insertion of less-difficult malignant cases on subsequent breast cancer detection by breast imaging radiologists.

Methods: The research comprises two studies. Study 1: 8 radiologists read 2 sets of images each consisting of 40 mammographic cases. Set A contained four abnormal cases, and Set B contained six abnormal cases, including two priming cases (less difficult malignancies) placed at intervals of three and five subsequent cases before a subtle cancer. Study 2: 16 radiologists read a third condition of the same cases, known as Set C, containing six abnormal cases and two priming cases immediately preceding the subtle cancer cases. The readers were asked to localize malignancies and give confidence ratings on decisions.

Results: Although not significant, a decrease in performance was observed in Set B compared with in Set A.

There was a significant increase in the receiver operating characteristic (ROC) area under the curve ($z = -2.532$; $p = 0.0114$) and location sensitivity ($z = -2.128$; $p = 0.0333$) between the first and second halves of Set A and a marginal improvement in jackknife free-response ROC figure of merit ($z = -1.89$; $p = 0.0587$) between the first and second halves of Set B. In Study 2, Set C yielded no significant differences between the two halves of the study.

Conclusion: Overall findings show no evidence that priming with lower difficulty malignant cases affects the detection of higher difficulty cancers; however, performance may decrease with priming.

Advances in knowledge: This research suggests that inserting additional malignant cases in screening mammography sets as an audit tool may potentially lead to a decrease in performance of experienced breast radiologists.

The effectiveness of screening mammography in detecting breast cancer relies heavily on accurate observer performance;¹ however, miss rates of 10–30% have been reported.^{2,3} The challenge for radiologists is to gain comprehensive experience in the appearance of abnormality and keep their recall rate within acceptable limits given the low incidence of the disease.^{4–6} It has been proposed that the insertion of “test malignant cases” into a “real life” clinical screening mammography set may increase cancer detection rates by artificially increasing the prevalence.⁷ However, by specifically increasing the prevalence of abnormality, there may be inadvertent effects on the detection of subsequent “real” cases.

Previous work on the impact on increasing abnormality prevalence on radiologic performance has provided diverse conclusions, with one study suggesting that varying prevalence

was unlikely to alter the accuracy of the radiologists,⁸ another suggesting increased diagnostic efficacy with increasing prevalence⁹ and a further arguing no significant effect.¹⁰ A study focusing on prevalence expectation, whilst showing no significant impact on reporting accuracy, did show that visual search, in terms of interpretation time and the number of visual fixations per image, was significantly changed when higher prevalence was expected.¹¹

A recent study of expert breast radiologists showed that they tend to miss a much higher percentage of malignant lesions in a setting closer to the actual low-prevalence level found in the clinic than under high prevalence test conditions.⁷ However, in that study, although the false-negative rate significantly increased in the clinic setting, the

false-positive rate decreased in the low-prevalence condition, although by a non-significant amount. Moreover, owing to the difficulties of performing such a study in the clinical setting, the same radiologists did not interpret the same inserted cases in both the low and high prevalence parts of the study, and as reader variability usually comprises the largest component of variance in a reader study, care should be taken when interpreting these results.

Research in cognitive psychology has shown that human visual perception of the elements in our environment is greatly dependent on what was previously seen.¹² For example, detection of a target is facilitated by having seen the target (or a lookalike) before, and this effect is called “perceptual priming”. Perceptual priming as a mechanism for improving cancer detection in low-prevalence contexts such as screening mammography is an under explored area of research. In this context, priming can be understood as a mechanism that “highlights” the presence of a given target by increasing the saliency of the stimulus.¹³ It is generally accepted in the cognitive psychology literature that priming can be used to direct visual attention towards an explicit target, and that observers become proficient at finding targets when salient prompts are employed.^{14,15} This has real relevance in clinical imaging, not just as an attempt to improve diagnostic efficacy but also to establish the impact of audit cases randomly inserted into a reading set to test radiologic performance. The hypothesis of the study is that the addition of malignant cases of lower difficulty (herein called “primers”) would increase the likelihood that the readers would detect subsequent malignant cases of higher difficulty presented at intervals between one and five cases after the primers.

METHODS AND MATERIALS

This research was conducted through two separate but related experiments, called Study 1 and Study 2. Institutional ethics approval was granted for these studies by the University of Sydney, Australia.

Participants

Study 1 was performed at a meeting of the American Board of Radiology in Louisville, KY, in May 2012. Eight radiologists specializing in breast imaging volunteered to read two sets of mammography cases (Sets A and B). The radiologists had a median 15 years’ experience and reported on average 5000 mammographic cases per year (range = 1500–10,000). Study 2 was undertaken at a meeting of the Royal Australian and New Zealand College of Radiology Breast Imaging Group in Darwin, Australia, in June 2013. A total of 15 breast radiologists volunteered to read one set of mammography cases (Set C). None of these radiologists were involved in Study 1. These radiologists had a median of 22 years of experience and reported on a median of 4000 mammographic cases per year (range = 1000–20,000).

Cases

For both studies, the cases were acquired from a research bank from BreastScreen, Australia, and were selected by the State Radiologist for BreastScreen, New South Wales, Australia, to represent a variety of breast density and solid malignant lesions. The images were full field digital mammograms acquired from six

units: Carestream (Rochester, NY), Phillips (Amsterdam, Netherlands), Agfa Healthcare (Mostrel, Belgium), GE Senograph® Essential and Senograph DS (Waukesha, WI), Fuji Medical Systems (Tokyo, Japan) and Sectra Medical Systems (Linköping, Sweden).

The malignant cases, represented by breast masses, were confirmed by a verified pathology biopsy and the normal images by a subsequent negative screen. Throughout this study, the malignant lesions are classified as either of “lower difficulty” (both screening radiologists identified the lesion when first read in the actual clinical practice) or of “higher difficulty” (the malignancy was originally missed by one screening radiologist and corrected by positive localization by a third radiologist, the arbitrator, which is the current practice in Australia). A total of six single lesion malignant cases were used throughout these studies, with the largest mass size being 75 pixels in radius; some lesions had microcalcifications within the mass boundaries.

Image display

The images for both studies were displayed on two five-megapixel medical diagnostic monochrome liquid crystal display EIZO® GS510 (EIZO, Hakusan, Japan) monitors and were de-identified. A Sectra workstation was used (Sectra Imtec AB, Linköping, Sweden) to drive the monitors, and the median resolution for the images was 4740×3540 pixels (maximum 5928×4728 pixels and minimum 2294×1914 pixels). The ambient lighting conditions were controlled to approximately 20 Lux and measured using a calibrated photometer (Model Konica Minolta CL-200, Ramsey, MN). The monitors were calibrated using the parameters of digital imaging and communications in medicine part 14 standard using an EIZO UX1 sensor and quality control software.

Experimental design

In Study 1, two test sets of cases (Sets A and B), each of 40 mammographic cases, were employed, with each case comprising two craniocaudal (CC) and two mediolateral oblique (MLO) images representing both breasts. The cases in Set A and B were the same except for the number of malignant cases that were present. Set A had four malignant cases of “higher” difficulty, and these were at case locations #9, #14, #25 and #37. Set B had an additional two malignant cases of lower difficulty replacing two malignancy-free cases at positions #20 and #34, as shown in Figure 1 (all other images were the same and in the same location for both sets). These additional inserted cases were positioned at these specific locations, preceding the “higher difficulty” lesions, so as to investigate the effect that a likely detected cancer would have on radiologists’ performance. For Study 2, a third set (Set C) was developed and implemented at a subsequent meeting to investigate the immediacy of any effect, so now the locations of positive cases in the 40 case “Set C” test set were at #9, #14, #20, #21, #34 and #35, with #20 and #34 being of “lower difficulty” (*i.e.* the primers). The number of inserted cases was deliberately kept to a very low level so as to not increase the possibility of alerting readers to the objective of the study.

This arrangement of cases created for Sets A, B and C allowed for a first, unprimed half without inserted lower-difficulty cases

Legend

- Normal
- Abnormal (high difficulty)
- Abnormal (*primer*: low difficulty)

Case Order

1 Midpoint 40

1st Half **2nd Half**

9 14 25 37

20 34

21 35

Control

Sets containing primers

Primer

There was no time limit for the readings. A counterbalancing protocol was employed whereby reading Set A and Set B were alternated for each breast radiologist in Study 1. For Study 1, the radiologists performed the two separate sessions of image reading with an interval of 1 and 3 days between sittings. For

For both studies, observer performance was calculated using the following metrics: sensitivity, specificity, receiver operating characteristic (ROC) area under the curve (Az), jackknife free-response ROC (JAFROC) figure of merit (FOM) and location sensitivity. A true positive was scored as being a correct localization within a 100-pixel radius of the centre of the lesion. A false positive was defined as an incorrect localization of any non-lesion area, >100 pixels from the nearest possible true lesion or any mark on a normal case. A false negative was scored as being the lack of a mark within 100 pixels from the centre of a true lesion in an abnormal case. Cases #20 and #34 were omitted from the analysis as they were “primers”, and >95% of all readers correctly localized these lesions on all reads. The Mann–Whitney *U*-test was used to perform the comparisons. A *p*-value of <0.05 was used to determine statistical significance.

The mean, standard deviation, Mann–Whitney’s Z - and p -values for comparisons between Set A and Set B for ROC, IAFROC,

location sensitivity, sensitivity and specificity for all breast readers are shown in Table 1. Overall, no significant differences were recorded in any performance measure when Set A and Set B were compared, although some small non-significant decreases in performance were noted in the latter.

Tables 2 and 3 show the mean performance metrics within each half of Sets A and B. When we compare the performance on each half between Set A and Set B, we found no statistically significant differences, although a decline in sensitivity in the second half between Set A (mean = 1.000) and Set B (mean = 0.875) and in location sensitivity between Set A (mean = 0.938) and Set B (mean = 0.813) was noticed.

Observer performance was calculated within the sets, with comparisons made between performance in the first (Cases 1–19) and second (Cases 21–40) halves, respectively. There is a significant difference in the ROC Az ($z = -2.532$; $p = 0.0114$) between the first half (mean = 0.802) and second half (0.960) of Set A. Location sensitivity also showed a significant increase from 0.500 in the first half to 0.938 in the second half ($z = -2.128$; $p = 0.0333$) (Table 2), and there is a trend towards significance in the JAFROC FOM scores from 0.57 in the first half to 0.893 in the second half ($z = -1.842$; $p = 0.066$) (Table 2) when comparing Cases 1–19 and 21–40. Conducting the same comparison of halves in Set B (Table 3), again the JAFROC FOM scores indicate a trend towards significance from 0.57 in the first half to 0.829 in the second half and a greater number of readers may see a significant difference confirmed ($z = -1.890$; $p = 0.0587$).

Study 2

For Study 2 (Set C), the mean, standard deviation and the Mann–Whitney's Z - and p -values for ROC, sensitivity and specificity scores were calculated for the first half and second half comparisons and are shown in Table 4. Although no significant differences were found between the first and second halves of Set C, a trend was observed towards lower ROC Az in the second half (mean = 0.929) as compared with the first (mean = 0.952; $z = -1.748$; $p = 0.081$). Similarly, the JAFROC FOM was lower on the second half (0.847) than on the first half (0.884), but the difference was not significant ($z = -1.678$; $p = 0.093$).

DISCUSSION

This study enabled us to measure the impact of inserting lower difficulty malignant cases, as a means of perceptual priming upon the subsequent performance of expert breast radiologists in the detection of malignancies of higher difficulty. No evidence has been provided by this study to support the hypothesis that the addition of malignant cases of lower difficulty would increase the readers' ability to further detect subsequent malignant cases of higher difficulty at intervals between one and five cases. Some small decreases in performance were noted for between the first and second halves in Set B (JAFROC FOM, Table 3) and again in Set C (ROC Az, Table 4). Because this decrease only occurred in the two sets with the primers, it is possible that the presentation of the "lower difficulty" lesions prompted the radiologists to "detect cancers" even in the normal cases that followed the primer. This could potentially explain why the effect of inserting lower difficulty malignancy cases was stronger in Set B, where there were two or four normal cases between the primer and the next true cancer case, than in Set C, where there were only one or two normal cases in between. Research in cognitive psychology has suggested that priming has a more significant effect immediately after fixation of the primer (e.g., after looking at a red rose, the eyes are more attracted to either another red rose or to a likely object).^{16,17} However, we cannot eliminate a low power effect from these results, which will be addressed in future works by using greater numbers of normal and abnormal cases.

For Set A, there was a statistically significant increase in the performance of the radiologists in the second half of the test set, where we observed an increase in location sensitivity without a change in specificity. One possible explanation for this increase may be that the radiologists were becoming familiar with the experimental setting and were developing a sense of how likely it was that they would find cancers within the test set. In Set B, where the primer cases were present, no such improvements were noticed, although the JAFROC FOM increased from the first half (mean = 0.570) to the second half (mean = 0.829), and this was marginally significant ($p = 0.0587$). This was accompanied by a small non-significant improvement in specificity (first half mean = 0.559; second half mean = 0.640; $p = 0.183$). Taken together, these results suggest that the primers may have had some effect on the breast radiologists. Unfortunately, we

Table 1. Study 1: performance metrics for all eight readers in Set A (no priming) and Set B (with priming)

Performance metrics	Set A (no priming), mean (SD)	Set B (with priming), mean (SD)	Z	p -value
Receiver operating characteristic area under the curve	0.884 (0.065)	0.845 (0.121)	-0.684	0.4940
Sensitivity	0.938 (0.166)	0.875 (0.189)	-0.645	0.5186
Specificity	0.634 (0.100)	0.599 (0.110)	-0.688	0.4920
Jackknife free-response receiver operator characteristics figure of merit	0.712 (0.162)	0.701 (0.163)	0.000	0.9990
Location sensitivity	0.719 (0.248)	0.688 (0.248)	-0.226	0.8120

SD, standard deviation.

Table 2. Study 1: comparison between the two halves of Set A only.

Performance metrics	Set A first half, mean (SD)	Set A second half, mean (SD)	Z	p-value
Receiver operating characteristic area under the curve	0.802 (0.160)	0.960 (0.038)	-2.532	0.011 ^a
Sensitivity	0.875 (0.231)	1.00 (0.000)	-1.464	0.143
Specificity	0.640 (0.164)	0.647 (0.109)	-0.160	0.873
Jackknife free-response receiver operator characteristics figure of merit	0.570 (0.268)	0.839 (0.141)	-1.842	0.066
Location sensitivity	0.500 (0.463)	0.938 (0.177.000)	-2.138	0.033 ^a

SD, standard deviation.

No priming cases were present in this set.

^aIndicates significance.

cannot assess which effect this may have been, because of the potential impact of memory effect, which cannot be disregarded in this experimental setting. Owing to the short time lapse between the readings, it is possible that memory traces from the first reading influenced the second reading.

Our study does not concur with recent literature by Evans et al⁷ and others,^{13,14,18} as we have not found a sensitization, or priming effect, to the presence of breast cancer by inserting cases to prime the reader. Our study does support the previous work of Gur et al¹⁰ in documenting a lack of correlation in performance change through altered prevalence, although this may be influenced by the laboratory conditions and the subtle change in abnormal case numbers.¹⁵ Also a higher number of abnormal cases were used than what the readers would expect to find in their usual case sets. However, this was necessary owing to the excessive number of cases that the reader would need to interpret to mimic clinical screen reading prevalence. Also this study was laboratory based, and it is necessary to explore if conclusions provided here are relevant in a clinic-based study. Furthermore, all participating radiologists were highly experienced, so their responses to the primers may not reflect those of the general population of radiologists.

The number of cases used was lower than what would be ideal. However, the number of cases used in this preliminary work was to facilitate the engagement of a good number of radiologists,

which would have been difficult in one sitting if too many cases were presented in the experimental reading settings used here. Nonetheless, future work will address this issue by using a greater image and case number.

This work was an exploratory study to determine if any perceptual priming could be observed by the insertion of relatively easy positive cases. Whilst no direct evidence was shown for priming, some small changes found here need further study. The current work does not address a number of important issues that would need investigation if this type of approach was ever to be used in clinical imaging, such as the timing of the cases and the practical difficulty in anticipating when abnormal cases occur, technical challenges with inserting cases seamlessly into a radiologist's workload, and the type of lesion that should be used as a primer. All these are currently being considered by our group and future work will focus on these issues, the removal of any memory effect and the impact of priming on less experienced radiologists. The number of both normal and abnormal cases included within each test set will be increased.

CONCLUSION

In summary, this preliminary study indicates that inserting lesser difficulty malignant cases into a clinical set does not have a positive effect on expert readers' detection of subsequent breast cancers and may have a detrimental effect, although our study was unable to show this change to a significant level. These

Table 3. Study 1: comparison between the two halves of Set B (with priming). Only primer Cases #20 and #34 were excluded from analysis

Performance metrics	Set B first half, mean (SD)	Set B second half, mean (SD)	Z	p-value
Receiver operating characteristic area under the curve	0.801 (0.152)	0.879 (0.151)	-1.214	0.225
Sensitivity	0.875 (0.231)	0.875 (0.231)	0.000	0.999
Specificity	0.599 (0.113)	0.640 (0.111)	-1.330	0.183
Jackknife free-response receiver operator characteristics figure of merit	0.570 (0.258)	0.829 (0.140)	-1.890	0.059
Location sensitivity	0.563 (0.417)	0.813 (0.259)	-1.273	0.203

SD, standard deviation.

Table 4. Study 2: comparison between the two halves of Set C (with priming, shorter interval).

Performance metrics	Set C first half, mean (SD)	Set C second half, mean (SD)	Z	p-value
Receiver operating characteristic area under the curve	0.942 (0.094)	0.929 (0.080)	−1.748	0.081
Sensitivity	0.967 (0.129)	0.967 (0.129)	0.000	0.999
Specificity	0.686 (0.194)	0.701 (0.121)	−0.042	0.967
Jackknife free-response receiver operator characteristics figure of merit	0.884 (0.206)	0.847 (0.182)	−1.678	0.933
Location sensitivity	0.883 (0.229)	0.833 (0.278)	−0.642	0.644

SD, standard deviation.

Primer Cases #20 and #34 were omitted from the analysis.

changes, following the insertion of priming cases, justify the need for a larger investigation designed to consider a number of the issues raised by this article.

FUNDING

This study was supported by the MIOPeG Carestream Healthcare funding source to fund travel and accommodation expenses.

ACKNOWLEDGMENTS

The authors would like to acknowledge the trustees, oral examiners and staff of the American Board of Radiology for their assistance and participation. Their thanks to the staff and radiologists of the Breast Imaging Group, Breast Interest Group of the Royal Australian and New Zealand College of Radiology for their involvement. Their thanks to Sectra staff for the loan of monitors at the Darwin Conference Centre.

REFERENCES

- Kopans DB, Smith RA, Duffy SW. Mammographic screening and “overdiagnosis”. *Radiology* 2011; **260**: 616–20.
- Bird RE, Wallace TW, Yankaskas BC. Analysis of cancers missed at screening mammography. *Radiology* 1992; **184**: 613. doi: [10.1148/radiology.184.3.1509041](https://doi.org/10.1148/radiology.184.3.1509041)
- Majid AS, de Paredes ES, Doherty RD, Sharma NR, Salvador X. Missed breast carcinoma: pitfalls and pearls. *Radiographics* 2003; **23**: 881. doi: [10.1148/rg.234025083](https://doi.org/10.1148/rg.234025083)
- Gur D, Sumkin JH, Rockette HE, Ganott M, Hakim C, Hardesty L, et al. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. *J Natl Cancer Inst* 2004; **96**: 185–90.
- Pisano ED, Gatsonis C, Hendrick E, Yaffe M, Baum JK, Acharyya S, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. *N Engl J Med* 2005; **353**: 1773–83. doi: [10.1056/NEJMoa052911](https://doi.org/10.1056/NEJMoa052911)
- Sickles EA, Miglioretti DL, Ballard-Barbash R, Geller BM, Leung JWT, Rosenberg RD, et al. Performance benchmarks for diagnostic mammography. *Radiology* 2005; **235**: 775–90. doi: [10.1148/radiol.2353040738](https://doi.org/10.1148/radiol.2353040738)
- Evans KK, Birdwell RL, Wolfe JM. If you don’t find it often, you often don’t find it: why some cancers are missed in breast cancer screening. *PLoS One* 2013; **8**: e64366. doi: [10.1371/journal.pone.0064366](https://doi.org/10.1371/journal.pone.0064366)
- Gur D, Rockette HE, Warfel T, Lacomis JM, Fuhrman CR. From the laboratory to the clinic: the “prevalence effect”. *Acad Radiol* 2003; **10**: 1324–6.
- Eggin TKP, Feinstein AR. Context bias: a problem in diagnostic radiology. *JAMA* 1996; **276**: 1752–5.
- Gur D, Rockette HE, Armfield DR, Blachar A, Bogan JK, Brancatelli G, et al. Prevalence effect in a laboratory environment. *Radiology* 2003; **228**: 10–14. doi: [10.1148/radiol.2281020709](https://doi.org/10.1148/radiol.2281020709)
- Reed WM, Ryan J, McEntee MF, Evanoff M, Brennan PC. The effect of abnormality-prevalence expectation on expert observer performance and visual search. *Radiology* 2011; **258**: 938–43. doi: [10.1148/radiol.10101090](https://doi.org/10.1148/radiol.10101090)
- Kristjansson A, Campana G. Where perception meets memory: a review of repetition priming in visual search tasks. *Atten Percept Psychophys* 2010; **72**: 5–18.
- Theeuwes J, Van der Burg E. Priming makes a stimulus more salient. *J Vis* 2013; **13**: 1–11. doi: [10.1167/13.3.21](https://doi.org/10.1167/13.3.21)
- Brascamp JW, Blake R, Kristjansson Á. Deciding where to attend: priming of pop-out drives target selection. *J Exp Psychol Hum Percept Perform* 2011; **37**: 1700–7. doi: [10.1037/a0025636](https://doi.org/10.1037/a0025636)
- Wolfe JM, Horowitz TS, Van Wert MJ, Kenner NM, Place SS, Kibbi N. Low target prevalence is a stubborn source of errors in visual search tasks. *J Exp Psychol Gen* 2007; **136**: 623–38. doi: [10.1037/0096-3445.136.4.623](https://doi.org/10.1037/0096-3445.136.4.623)
- Gur D, Bandos A, Fuhrman CR, Klym AH, King JL, Rockette HE. The prevalence effect in laboratory environment: changing the confidence ratings. *Acad Radiol* 2007; **14**: 49–53.
- Kristjansson A. “I know what you did on the last trial”—a selective review of research on priming in visual search. *Front Biosci* 2008; **13**: 1171–81.
- Wolfe JM, Horowitz TS, Kenner NM. Cognitive psychology: rare items often missed in visual searches. *Nature* 2005; **435**: 439–40. doi: [10.1038/435439a](https://doi.org/10.1038/435439a)