European Journal of Radiology Open 10 (2023) 100498

Contents lists available at ScienceDirect

European Journal of Radiology Open

journal homepage: www.elsevier.com/locate/ejro

Original article

# Reducing ' probably benign ' assessments in normal mammograms: The role of radiologist experience

Mohammad A. Rawashdeh [a,b,*], Patrick C. Brennan [c]

[a] Faculty of Health Sciences, Gulf Medical University, Ajman, United Arab Emirates
[b] Faculty of Applied Medical Sciences, Jordan University of Science and Technology, Irbid 222110, Jordan
[c] Medical Image Optimisation and Perception Group (MIOPeG), Sydney School of Health Sciences, Faculty of Medicine and Health, The University of Sydney, Sydney, New South Wales, Australia

ARTICLE INFO

*Keywords:*
Mammography
Radiologist performance
Probably benign

ABSTRACT

*Rationale and objectives:* to investigate the relationship between radiologists' experience in reporting mammograms, their caseloads, and the classification of category '3′ or 'Probably Benign' on normal mammograms.
*Materials and Methods:* A total of 92 board-certified radiologists participated. Self-reported parameters related to experience, including age, years since qualifying as a radiologist, years of experience reading mammograms, number of mammograms read per year, and hours spent reading mammograms per week, were documented. To assess the radiologists' accuracy, "Probably Benign fractions" was calculated by dividing the number of "Probably Benign findings" given by each radiologist in the normal cases by the total number of normal cases Probably Benign fractions were correlated with various factors, such as the radiologists' experience.
*Results:* The results of the statistical analysis revealed a significant negative correlation between radiologist experience and 'Probably Benign' fractions for normal images. Specifically, for normal cases, the number of mammograms read per year ($r = -0.29$, $P = 0.006$) and the number of mammograms read over the radiologist's lifetime ($r = -0.21$, $P = 0.049$) were both negatively correlated with 'Probably Benign' fractions
*Conclusion:* The results indicate that a relationship exists between increased reading volumes and reduced assessments of 'Probably Benign' in normal mammograms. The implications of these findings extend to the effectiveness of screening programs and the recall rates.

## 1. Introduction

Screening mammography has been widely acknowledged as the preferred diagnostic modality for detecting breast cancer [1]. The American College of Radiology (ACR) created the BI-RADS (Breast Imaging-Reporting and Data System) as a means of evaluating risk and maintaining quality in breast imaging. This system was developed with the goal of improving accuracy in mammogram assessment, reducing interpretive errors, and enhancing care for women experiencing symptomatic breast conditions. It provides a standardized vocabulary and reporting structure for breast imaging, including mammography, ultrasound, and MRI. This article is based on the 5th edition of BI-RADS, which was published in 2013 [2]. This lexicon includes the following categories: Category 1 denotes no significant abnormality and requires no further imaging; Category 2 indicates benign findings and necessitates no further imaging; Category 3 corresponds to Probably Benign findings that require additional investigation, typically short-interval (6-month) follow-up; Category 4 represents suspicious findings of malignancy that necessitate further investigation and possibly excisional biopsy; Category 5 highly suggestive of malignancy signifies malignant findings that mandate additional investigation, even if non-excision (percutaneous) sampling indicates benignity; Category 0, indicating incomplete assessment, and Category 6, indicating biopsy-proven malignancy. These categories are analogous to those recommended by The Royal College of Radiologists Breast Group in the UK [3]. However, the ACR (BI-RADS), which is frequently used in North America and parts of Europe. Moreover, BI-RADS cannot be directly applied to the Australian context because it recommends biopsy rather than short-term follow-up for Category 3 or equivocal findings [2–4].

The mid-point, referred to as 'category 3′ or 'Probably Benign',

*Abbreviations:* ACR, American College of Radiology; BI-RADS, Breast Imaging-Reporting and Data System.

* Corresponding author at: Faculty of Health Sciences, Gulf Medical University, Ajman, United Arab Emirates.
*E-mail address:* dr.rawashdeh@gmu.ac.ae (M.A. Rawashdeh).

https://doi.org/10.1016/j.ejro.2023.100498
Received 11 April 2023; Received in revised form 7 June 2023; Accepted 9 June 2023
2352-0477/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

denotes mammographic findings that have a low probability of malignancy a 2 % or less [4–6] and require typically short-interval follow -up and biopsies for suspicious lesions. However, a small percentage of findings designated as probably benign are later upgraded to suspicious and require biopsy. The use of BI-RADS 3 is challenging, as there is significant variability among breast imagers in their assessments of these findings. Formal instruction has been shown to enhance the precision of BI-RADS assessments, with BI-RADS 1 and 2 are straightforward, while BI-RADS 4 and 5 are indicative of suspicious or highly suspicious results. Misuse of this category is likely to result in the unnecessary recall of more women, leading to emotional distress such as anxiety, stress, and pain [5], as well as increased screening costs [7–11] due to the possible need for fine needle aspiration (FNA) cytology or core biopsy. Previous studies investigated the link between radiologists' characteristics and diagnostic accuracy or performance in the mammography [12–21]. Others analyzed the relationship between cancer detection and mammographic density, lesion locations, or image features [22–27]. To our knowledge, this is the first study to investigate the impact of radiologists' experiences on assigning the 'Probably Benign' category of mammograms.

The present study aims to explore how demographic factors and experience may reduce the use of 'Probably Benign' classification on normal mammograms and alleviate the negative impacts associated with this practice, such as falsely recalling women and the unnecessary performance of biopsies. Specifically, the study will investigate the relationship between radiologists' experience in reporting mammograms, their caseloads, and the classification of category '3′ or 'Probably Benign' on normal mammograms. By addressing this research question, the study aims to contribute to the optimization of mammogram reporting practices and improve the efficiency and quality of breast cancer screening programs.

## 2. Material and methods

### 2.1. Image set

Prior to conducting the study, approval was obtained from the institutional ethics review board. The data were initially collected in 2019 for a different objective and were subsequently reanalyzed in 2023 to align with the current objective. The test set comprised a total of 60 mammogram cases, each of which consisted of four images, including caudal cranial (CC) and mediolateral oblique (MLO) projections for each breast, resulting in a total of 240 images. Among these cases, twenty were confirmed to have biopsy-proven cancer, with four of these cases containing multiple lesions. The remaining forty images were confirmed to be normal through follow-up mammograms conducted two years later. The normal cases contained incidental benign findings, such as calcified duct ectasia, calcified oil cysts, benign calcified fibroadenoma, and intramammary lymph nodes.

### 2.2. Radiologist's experience details

In this study, a total of 92 board-certified radiologists participated. Self-reported parameters related to experience, including age, years since qualifying as a radiologist, years of experience reading mammograms, number of mammograms read per year, and hours spent reading mammograms per week, were documented. To calculate the experience level of each radiologist, the number of mammograms read over their lifetime was independently calculated by multiplying the number of years reading mammograms by the number of mammograms read per year. A summary of the radiologists' details can be found in Table 1.

### 2.3. Test environment

The radiologists interpreted the images in one of two rooms, each measuring 60 $m^2$ and 90 $m^2$, with walls painted in light gray and brown matte colors to minimize specular reflection. To ensure a consistent lighting environment, a calibrated photometer (Model Konica Minolta CL-200, Ramsey, NJ) was used to assess ambient light, which was maintained between 12 and 20 lux. The workstations used for the study were equipped with monitors of varying sizes and models, video cards, and calibration, as described in Table 2.

**Table 1**
Details on the 92 participating radiologists, with interquartile rangers.

| Parameter | Median | First quartile | Third quartile | Min | Max |
|---|---|---|---|---|---|
| **Age (y)** | 52.5 | 44.5 | 56.75 | 31 | 75 |
| **Years since qualification** | 13.5 | 8.5 | 20 | 1 | 42 |
| **Years reading mammograms** | 10 | 4.25 | 16 | 1 | 28 |
| **Mammograms read per year** | 2500 | 1213 | 5000 | 250 | 15,000 |
| **Hours reading mammograms per week** | 10 | 5 | 19.5 | 1 | 40 |
| **Mammograms read over lifetime*** | 24,000 | 8100 | 49,500 | 500 | 300,000 |

* The number of years reading mammograms multiplied by the number of mammograms read per year

### 2.4. Study description

Radiologists were asked to localize and assess breast abnormalities according to BIRADS assessment categories. To facilitate this task, the software platform used was the Breast Reader Assessment Strategy (BREAST), which enabled the reading of digital images, determination of lesion location, and provision of an assessment category for breast lesions. The assessment categorization involved giving any perceived lesion a score of 2 (benign), 3 (Probably Benign), 4 (suspicious) and 5 (malignant). No information regarding the number of abnormal or normal cases was provided, and all radiologists were given an explanation of the test software prior to commencing the test. There was no time limit for the assessment of images, and radiologists were able to freely access post-processing tools such as panning, zooming, and windowing. After arriving at a decision, radiologists used a mouse-controlled cursor to locate any perceived lesion on a laptop that presented the same image as the one displayed on the high-resolution monitors. If the decision about the case was that it was "normal", radiologists could simply click on "next case" and the category score 1 (negative) would automatically be assigned for that case.

**Table 2**
Specifications of the workstations.

| Parameters | Workstations | |
|---|---|---|
| **Monitor** | Sectra, Linköpking, Sweden MFGD 5621; Barco, Kortrijk, Belgium | Hologic, Bedford, Mass RadiForce G51; Eizo, Ishikawa, Japan |
| **Monitor size** | 5 megapixel | 5 megapixel |
| **Video Card** | BarcoMed 5MP2FH | Matrox MED5MP-DVI; Dorval, Quebec, Canada |
| **Calibration** | Digital Imaging and Communication in Medicine | Digital Imaging and Communication in Medicine |
| **Standard display function** | Gray-scale | Gray-scale |
| **Minimum luminance** | 1.3 cd/$m^2$ | 1.3 cd/$m^2$ |
| Maximum luminance | 5% of 475 cd/m2 | 5% of 475 cd/$m^2$ |
| **Contrast Ratio** | 365:1 | 365:1 |
| **Number of workstations** | 2 | 2 |

### 2.5. Data statistical analysis

To determine the radiologist characteristics that could lead to increased or decreased usage of the Probably Benign score 3, referred to as 'Probably Benign', the study's first step involved calculating the 'Probably Benign' fractions for each radiologist. This fraction was calculated by dividing the number of normal cases with an 'Probably Benign' score given by each radiologist by the total number of normal cases, which consisted of 40 normal cases.

The "Probably Benign" fractions for each group of cases were then independently correlated with each of the radiologist experience parameters listed in Table 4 using non-parametric Spearman techniques. Additionally, the radiologists were categorized based on their number of readings per year into several categories: $\leq 999$, $> 999$, 1000–1999, $\leq 1999$, $> 1999$, 2000–4999, $\leq 4999$, and $> 4999$. Experience parameters were then calculated for each of these groups and correlated with the "Probably Benign" fractions as described above. Further analysis involved a stepwise linear regression to predict the independent impact of significant radiologist experience parameters on the 'Probably Benign fractions'.

All statistical analyses were performed using the software IBM SPSS Statistics (version 22.0, for MAC; SPSS). Results were considered statistically significant when the P value was $\leq 0.05$.

## 3. Results

The results of the statistical analysis revealed a weak significant negative correlation between radiologist experience and 'Probably Benign' fractions for normal images. Specifically, for normal cases, the number of mammograms read per year ($r = -0.29$, $P = 0.006$) and the number of mammograms read over the radiologist's lifetime ($r = -0.21$, $P = 0.049$) were both negatively correlated with 'Probably Benign' fractions (Table 3).

For the group of radiologists reading more than 1000 mammograms per year, the total number of mammograms per year was statistically a weak negative correlated with 'Probably Benign fractions' for normal cases only ($r = -0.2331$ and $P = 0.0351$), as detailed on Table 4.

The stepwise linear regression analysis revealed significant predictors for normal images, specifically the number of mammograms read per year ($F = 9.622$, $p = 0.003$). These results suggest that both factors are influential in predicting outcomes for normal images, and may have implications for enhancing the precision and effectiveness of mammogram interpretation in clinical settings.

**Regression formula** $=0{\cdot}132–5{\cdot}056\times10^{-6} \times$ (mammograms/year)

## 4. Discussion

Studying the "Probably Benign" category in mammogram readings is crucial as it is associated with a low malignancy percentage of less than 2 % and requires short-term follow-up. False recall of women due to Probably Benign readings can result in high healthcare costs and emotional distress. Therefore, it is important to reduce inaccurate uses of this assessment category in mammograms. This study provides valuable information regarding which radiologists' experiences and caseloads may impact the BIRADS 3 or Probably Benign assessment of mammograms [1–5]. To the best of our knowledge, this is the first study to investigate this issue, and the findings may be useful in improving the accuracy and effectiveness of mammogram readings, ultimately benefiting patient outcomes.

The present study's findings indicate a statistically significant negative correlation between the number of mammograms read per year and Probably Benign assessment of normal cases. These results suggest that radiologists with higher volume of readings were less likely to assign Probably Benign scores to normal breasts, indicating that volume of reading plays a significant role in the accuracy of breast cancer diagnosis. Additionally, the study reveals a negative correlation between radiologists who read over 1000 mammograms per year and Probably Benign assessment. In summary, these results suggest that radiologists can reduce Probably Benign assessment of mammograms in normal cases by increasing their annual readings of mammograms, with particular emphasis on reading over 1000 mammograms per year. While no previous studies have examined the correlation between radiologists' experience and Probably Benign assessment, prior research has established a link between radiologists' experience and mammogram assessment performance and accuracy. Earlier studies have demonstrated that higher volumes of mammograms read per year and increased weekly hours spent reading mammograms are associated with higher reader performance. Additionally, radiologists who read less than 1000 mammograms annually perform worse than those who read more than 1000 mammograms annually. The current study's findings support these earlier studies, as they suggest that higher volumes of mammograms read per year and increased weekly hours spent reading mammograms can reduce errors in mammogram assessment [13–18].

Radiologists may encounter challenges in accurately evaluating abnormal mammograms due to the wide range of breast lesions and their various pathological features, including shapes, locations, margins, and sizes. Additionally, normal mammograms can also present challenges due to different breast densities. Therefore, there is a need for educational strategies to improve radiologists' experience and facilitate their observations of a large number of mammography cases. One example of an educational program is Detected X platform, which is designed to assist radiologists in correctly identifying pathological and non-pathological breast findings through the analysis of a large image database of abnormal and normal cases. Educational programs like Detected X have been shown to be effective in various medical domains, including auscultation, electrocardiogram analysis, and surgical Simulations. Existing platforms provide avenues for continuous auditing and training, as well as offering feedback to help radiologists improve early detection of breast cancer through mammography, including task-specific feedback. Task-specific feedback has been found to enhance performance in other medical fields, as noted by Choi et al. [28] who reported that feedback directed by hospitals improved emergency medical services performance. Previous studies have also shown that feedback from patients and colleagues can enhance physicians' clinical skills [29]. Therefore, the high frequency of feedback from physicians and the repeated interpretation of mammograms may explain why the number of cases read per year was found to be linked to fewer assessments of 'Probably Benign' in normal mammograms, rather than experience, in this study. Further research is necessary to fully explore these relationships and determine the most effective approaches to

**Table 3**
Shows the correlations between the 'Probably Benign fractions' of normal cases and radiologists' experiences.

| Truth | | Age (y) | Years since qualification | Years reading mammograms | Mammograms read per year | Hours per week reading mammograms | Mammograms read over lifetime |
|---|---|---|---|---|---|---|---|
| Normal images | **r** | 0.016 | -0.154 | -0.089 | **-0.286** | -0.203 | **-0.206** |
| | **P** | 0.881 | 0.141 | 0.398 | **0.006** * | 0.052 | **0.049** * |

*Indicates that there is a significant difference ($P < 0.05$)

**Table 4**
Shows r and P values for the correlation of probably benign fractions with number of mammograms per year.

| Truth | | 0–999 | > 999 | 1000 – 1999 | 0–1999 | > 1999 | 2,000-4999 | 0–4999 | > 4999 |
|---|---|---|---|---|---|---|---|---|---|
| Normal | Number | 10 | 82 | 17 | 27 | 65 | 38 | 65 | 27 |
| | Mean | 0.1460 | 0.1099 | 0.1253 | 0.1330 | 0.1058 | 0.1158 | 0.1229 | 0.0918 |
| | r | -0.1077 | -0.2331 | 0.1325 | -0.2016 | -0.1821 | 0.1671 | -0.0953 | -0.1132 |
| | P | 0.7249 | 0.0351 | 0.6084 | 0.3132 | 0.1465 | 0.3159 | 0.4499 | 0.5741 |

training and feedback for radiologists in the field of mammography. Therefore, educational intervention programs may be an effective strategy for radiologists to evaluate mammograms more accurately and reduce errors.

This study had several limitations that should be acknowledged. Firstly, the number of cases analyzed and radiologists participating in the study were relatively small, which may limit the generalizability of the findings. Secondly, the cases presented did not include the clinical histories of the patients or previous images to compare with the ones shown in the study, which could have affected the radiologists' evaluations. Finally, the reading over a lifetime was not self-reported data collected but an estimated calculation, which may have introduced some degree of inaccuracy in the results. These limitations should be considered when interpreting the findings of this study and should be addressed in future research to improve the validity and reliability of the results.

In conclusion, the 'Probably Benign' reporting of mammograms can have serious negative impacts, including increased health costs, unnecessary screening and procedures, and distress in women who are falsely recalled. The findings of this study suggest that higher volumes of reading is associated with fewer assessments of 'Probably Benign' in normal mammograms. Future research should further explore the efficacy of educational programs and other interventions to improve the accuracy of mammogram assessments and reduce the negative consequences of 'Probably Benign' reporting.

## Funding

This publication was not supported by any financial grants or funding, and the authors have no conflicts of interest to disclose.

## IRB was obtained

IRB was obtained from Jordan University of Sciences and Technology.

## CRediT authorship contribution statement

**Mohammad Rawashdeh:** Conceptualization, Methodology, Data curation, Writing- original draft preparation. **Patrick Brennan:** Conceptualization, Methodology, Data curation, Writing- review & editing,

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] J.G. Elmore, et al., Screening for breast cancer, JAMA 293 (10) (2005) 1245–1256.
[2] American College of Radiology. Breast imaging reporting and data system (BI-RADS) 5. Reston: American College of Radiology; 2013. 3.
[3] A.J. Maxwell, et al., The Royal College of Radiologists Breast Group breast imaging classification, Clin. Radiol. 64 (6) (2009) 624–627.
[4] Radiology, A.Co, A.C.R. Bi-Rads: Breast Imaging System: Breast Imaging Atlas: Mammography, Breast Ultrasound, Magnetic Resonance Imaging. 2003: American College of Radiology.
[5] R. Rosenberg, M. Linver, Use of BI-RADS 3–probably benign category in the American college of radiology imaging network digital mammographic imaging screening trial: Baum JK, Hanna LG, Acharyya S, et al. (Cambridge Health Alliance, MA; Brown Univ, Providence, RI; et al.) Radiology 260: 61-67, Radiology 23 (2) (2012) 147–148.
[6] R.G. Barr, et al., Probably benign lesions at screening breast US in a population with elevated risk: prevalence and rate of malignancy in the ACRIN 6666 trial, Radiology 269 (3) (2013) 701–712.
[7] J. Brodersen, V.D. Siersma, Long-term psychosocial consequences of false-positive screening mammography, Ann. Fam. Med. 11 (2) (2013) 106–115.
[8] J. Cockburn, et al., Psychological consequences of screening mammography, J. Med. Screen. 1 (1) (1994) 7.
[9] J.G. Elmore, et al., Ten-year risk of false positive screening mammograms and clinical breast examinations, N. Engl. J. Med. 338 (16) (1998) 1089–1096.
[10] S. Jha, Overdiagnosis versus overtreatment: a false dichotomy, Radiology 270 (2) (2014) (628-628).
[11] A.N. Tosteson, et al., Consequences of false-positive screening mammograms, JAMA Intern. Med. 174 (6) (2014) 954–961.
[12] L. Esserman, et al., Improving the accuracy of mammography: volume and outcome relationships, J. Natl. Cancer Inst. 94 (5) (2002) 369–375.
[13] D.L. Miglioretti, et al., Radiologist characteristics associated with interpretive performance of diagnostic mammography, J. Natl. Cancer Inst. 99 (24) (2007) 1854–1863.
[14] Rawashdeh, M.A., et al. Experience in reading digital images may decrease observer accuracy in mammography. in SPIE Medical Imaging. 2015. International Society for Optics and Photonics.
[15] D.S. Buist, et al., Effect of radiologists' diagnostic work-up volume on interpretive performance, Radiology 273 (2) (2014) 351–364.
[16] M.A. Rawashdeh, et al., Markers of good performance in mammography depend on number of annual readings, Radiology 269 (1) (2013) 61–67.
[17] W.E. Barlow, et al., Accuracy of screening mammography interpretation by characteristics of radiologists, J. Natl. Cancer Inst. 96 (24) (2004) 1840–1850.
[18] D.S. Buist, et al., Influence of annual interpretive volume on screening mammography performance in the United States, Radiology 259 (1) (2011) 72–84.
[19] J.G. Elmore, et al., Variability in interpretive performance at screening mammography and radiologists' characteristics associated with accuracy 1, Radiology 253 (3) (2009) 641–651.
[20] J.G. Elmore, C.K. Wells, D.H. Howard, Does diagnostic accuracy in mammography depend on radiologists' experience? J. Women'S. Health 7 (4) (1998) 443–449.
[21] W.M. Reed, et al., Malignancy detection in digital mammograms: important reader characteristics and required case numbers, Acad. Radiol. 17 (11) (2010) 1409–1413.
[22] M.A. Rawashdeh, et al., Quantitative measures confirm the inverse relationship between lesion spiculation and detection of breast masses, Acad. Radiol. 20 (5) (2013) 576–580.
[23] C. Mello-Thoms, et al., Understanding the role of correct lesion assessment in radiologists' reporting of breast cancer. Breast Imaging, Springer, 2014, pp. 341–347.
[24] Rawashdeh, M., et al. Measurement of breast lesion display luminance and overall image display luminance relative to optimum luminance for contrast perception. in SPIE Medical Imaging. 2011. International Society for Optics and Photonics.
[25] D.S.A. Mousa, et al., How mammographic breast density affects radiologists' visual search patterns, Acad. Radiol. 21 (11) (2014) 1386–1393.
[26] Al Mousa, D., et al. The impact of mammographic density and lesion location on detection. in SPIE Medical Imaging. 2013. International Society for Optics and Photonics.
[27] D.S.A. Mousa, et al., Mammographic density and cancer detection: does digital imaging challenge our current understanding? Acad. Radiol. 21 (11) (2014) 1377–1385.
[28] B. Choi, D. Tsai, C.G. McGillivray, C. Amedee, J.A. Sarafin, B. Silver, Hospital-directed feedback to emergency medical services improves prehospital performance. Stroke; a, J. Cereb. Circ. 45 (7) (2014) 2137–2140.
[29] J. Ferguson, J. Wakeling, P. Bowie, Factors influencing the effectiveness of multisource feedback in improving the professional practice of medical doctors: a systematic review, BMC Med. Educ. 14 (1) (2014) 1–12.