A SELF-DIRECTED LEARNING INTERVENTION FOR RADIOGRAPHERS RATING MAMMOGRAPHIC BREAST DENSITY

Abstract

Purpose: Subjective methods of mammographic breast density (MBD) assessment are prone to inter-reader variability. This work aims to assess the impact of a short self-directed experiential learning intervention on radiographers' reproducibility of MBD assessment.

Method: The study used two sets of images (test and learning intervention) containing left craniocaudal and left mediolateral oblique views. The test set had MBD ratings from Volpara™ and radiologists using the fourth edition Breast Imaging and Data Systems (BI-RADS®). Seven radiographers rated the MBD of the test set before and after a self-directed learning intervention using the percentage descriptors in the fourth edition BI-RADS® Atlas. The inter-reader agreement, agreement between radiographers and Volpara™ as well as radiologist, was assessed using a Weighted Kappa (κ_w).

Results: Overall, radiographers' inter-reader agreement (κ_w) was substantial (0.79; 95%CI: 0.70–0.87) before the intervention and almost perfect (0.84; 95%CI: 0.77–0.90) after the intervention. Before the intervention, radiographers demonstrated fair agreement with radiologists (0.24; 95%CI:-0.46–0.61) and Volpara™ (0.24; 95% CI: -0.41–0.59). A fair but slightly improved agreement was also observed between radiographers and radiologists (0.31; 95% CI: -0.33 - 0.64) as well as Volpara™ (0.28; 95% CI: -0.34- 0.61) after the intervention.

Conclusion: Findings demonstrate that a short duration self-directed experiential learning intervention reduces inter-reader differences in MBD classification, but has a negligible impact on improving the agreement between inexperienced and expert readers.

Keywords: Training; Mammography; BI-RADS®; Volumetric assessment; Volpara™; Inter-reader variability

Introduction

The proportion of dense tissue in a woman's breast is associated with the risk of developing breast cancer.[1, 2] Women with extremely dense breasts have a 4-6 fold higher risk of developing breast cancer compared to those with almost fatty breast.[1, 2] Breast density is also associated with traditional risk factors for breast cancer such as genetic, reproductive and lifestyle characteristics.[3, 4] The combination of breast density information with these risk factors has been shown to improve breast cancer risk prediction models.[5, 6] Also, breast density reduction over time is associated with a reduced risk of developing breast cancer, and intake of vegetables and Vitamin D is associated with lower breast density.[7] Therefore, clinical mammographic breast density (MBD) assessment may be relevant for generating cancer risk profile and tailoring interventions to reduce risk.

High MBD increases the risk of interval cancer (cancer detected within one year of a negative mammography screening outcome) and reduces the sensitivity of screening mammography.[1, 8] Interval cancer is linked to the high risk of cancer and the masking (camouflaging) effect due to tissue superimposition on two-dimensional (2D) mammography.[1, 3] The lower sensitivity of 2D mammography in dense breasts is due to masking effect and the similarity in mammographic appearance of dense tissue and cancer.[8] To mitigate the effects of MBD on cancer detection, imaging tools with 3D or pseudo-3D capabilities such as 3D ultrasound, digital breast tomosynthesis (DBT), and magnetic resonance imaging (MRI) have been introduced as adjuncts to mammography.[3] There is increasing advocacy for women to be notified of their breast density status, which has given rise to federal legislation in the USA mandating radiologists to produce breast density report.[9] This is because identifying women with dense breasts may facilitate informed decisions regarding appropriate imaging pathways that may improve early detection of

cancer in such women. Therefore, it is important that MBD assessment approaches are reproducible to ensure that clinical decisions regarding adjunctive imaging and screening frequency are made in a consistent fashion.

Of the methods developed for MBD assessment,[3, 10] subjective (visual) approaches such as BI-RADS® are the most commonly used clinically[3]. A major limitation of subjective MBD assessment is reader intra- and inter-reader variability,[3, 10] which has the potential to cause differences in clinical decision-making from MBD data. Volumetric methods have now been integrated into the MBD reporting framework to overcome human subjective variability and include Volpara™ and Quantra™.[3, 10] However, these tools are expensive and require additional computer servers to function,[3] and are therefore not an option for price sensitive healthcare systems. This, therefore, increases dependency on subjective approaches for clinical MBD classification. Previous studies have recommended training and retraining for MBD assessors to reduce BI-RADS® inter and intra-reader variability.[11, 12] Experiential learning is a long-established training methodology for developing competence in medicine, health and other fields.[13, 14] It involves observation of events or tasks, reflection and self-directed learning with the aim of developing competence suitable for practice.[13, 15] Literature shows that experiential learning which contain approaches such as training and mentorship, hands-on practice, self-directed learning and appropriate immediate feedback mechanisms with self-reflection can be effective in developing competence in students and novice practitioners.[14-16] No study has assessed the impact of a short training intervention with feedback on the reproducibility of subjective MBD assessment. Consequently, in our study we used a self-directed short computer-based experiential learning intervention with expert feedback in order to facilitate personal reflection and improve performance in the classification of MBD. If effective, short interventions

might be hosted online to provide e-Learning support for radiographers and others to develop and maintain competence in MBD assessment. In today's economically limited environment a learning approach such as this would be an important cost effective asset.

In the United Kingdom (UK), radiographers are heavily involved in screening mammography interpretation and imaging decision-making.[17] Although advanced practitioner and consultant radiographers' cancer detection performance in mammography interpretation is comparable to radiologists,[18] there is no data regarding their knowledge and inter-reader agreement in MBD assessment. Also, no study has assessed MBD agreement between radiographers and radiologists as well as other established MBD assessment approaches such as Volpara™. Consequently, the current work aims to assess inter-reader agreement of radiographers in MBD assessment and the impact of a self-directed, experiential learning exercise, with expert feedback on inter-reader agreement. It also aims to assess the agreement between radiographers and radiologists as well as Volpara™, and whether a self-directed, experiential learning exercise, with expert feedback, would improve radiographers' agreement with radiologists and Volpara™.

Methods

Institutional Review Board approval was granted for this study (IRB: 2013/448 and HREC protocol number 2014/905). The study was carried out in three phases; pre-test MBD assessment, an intervention, and post-test MBD assessment. The pre- and post-test images were viewed on a 5 Megapixel Barco (Kortrijk, Belgium) self-calibrating 21" display (MDNG – 5121). The ambient lighting was controlled between 20 and 30 lux. The learning intervention images were viewed on Dell (P2217H) 21.5" Full HD IPS LED displays.

Participants

Nine qualified radiographers were approached to participate. These radiographers held specialist postgraduate qualifications in mammography imaging practice. Also, they were nine months through an 11 month postgraduate university course which prepared them for the reporting of mammography images within the National Health Service Breast Screening Program (NHSBSP). Seven volunteered and gave consent to participate in the study. Participants completed a demographic questionnaire and then undertook the pre-test, the educational intervention, and the post-test MBD assessment.

Image selection

Two sets of images were selected, a test set and a learning intervention set. The same test set was used for the pre- and post- intervention MBD assessment. It comprised of 40 cases, each having three images: a left craniocaudal (LCC), left mediolateral oblique (LMLO), and a combination of LCC and LMLO presented together. These images were obtained from 40 women aged between 50 and 74 years and they were drawn from an American mammography screening database. All images were reported as normal (negative for cancer), with the women being returned for normal screening. All women consented to use of their images for research. The 40 cases had Volpara™ volumetric breast density (VDG) scores and the majority BI-RADS® report of 20 American Board of Radiology (ABR) examiners. The majority report in the current study refers to the consensus of at least 11 out of the 20 ABR radiologists. The learning intervention set comprised of 100 mammographic images containing CC views, with their MBD ratings by a group of three expert radiologists using the percentage descriptors in the fourth edition BI-RADS® Atlas (1: less than 25% dense; 2: 25%–50 %; 3: 50%–75%; 4: >75%). The intervention set provided

participants with immediate feedback on their rating of MBD, while participants received no feedback on their pre- or post-test performance.

Pre-intervention MBD assessment

The test set was used for the pre-intervention MBD assessment. All seven participants independently classified MBD of the cases into different percentages as described by the fourth edition BI-RADS® Atlas (1: less than 25% dense; 2: 25%–50 %; 3: 50%–75%; 4: >75%). Participants were told not to pan, zoom or change window settings as the study was not a lesion detection task. We recorded the MBD rating of each participant and generated a majority report. The majority report in this study represents the consensus of at least four of the seven participants involved in the reading. No feedback on performance was provided to participants at the end of the assessment. The readings of the participants were compared in pairs to assess their inter-reader agreement before the intervention. We also compared the majority report of participants to that of the ABR examiners (radiologists) as well as the VDGs assigned by Volpara™ to assess their agreement with these alternative measures prior to the intervention.

Self-directed, experiential learning exercise, with expert feedback (Learning Intervention)

The intervention set containing the CC views of 100 cases was presented to participants immediately after the pre-intervention assessment. We asked the participants to independently rate the percentage MBD on each image according to the percentage categories in fourth edition BI-RADS® descriptors and then to compare their rating to the expert radiologists rating. Thisfeedback from the ABR radiologists' readings to participants on their own rating was intended to help familiarise the participants with MBD appearances classified into different BI-RADS® categories by experts. Upon receiving feedback, the participants were asked to consider

the features on those mammograms that may have influenced ABR examiners ratings, note any differences and reflect on the features of the image and use the information for subsequent MBD rating. The learning intervention continued until the ratings of each participant matched that provided by the ABR examiners, indicating that participants were likely to be familiar with features that influenced radiologists' rating of MBD. All the intervention training took place in a single session of 2-3 hours. There was no time restriction for participants to conduct the learning intervention and to familiarise themselves with radiologists' MBD ratings.

Post-intervention MBD assessment

The post-intervention assessment was to test whether abilities developed during the self-directed learning experience from the intervention will reduce inter-reader variability and improve their agreement with the radiologists and Volpara^{TM} measurement. We presented the same test set three days after the pre-intervention MBD assessment. We recorded their post-intervention MBD rating and generated a majority report. Participants' ratings were compared in pairs to assess inter-reader agreement after the self-directed learning intervention. We also compared the post-intervention MBD rating of participants to that of radiologists as well as Volpara^{TM}.

Data Analysis

Statistical analysis was performed using the Statistical Package for Social Sciences (SPSS) Version 22. Agreement in MBD assessment between pairs of participants before and after the intervention was assessed using a Weighted Kappa (κ_w). A two-way mixed model was used to calculate average absolute agreement for all seven participants. This model was chosen because it selects cases randomly, nesting the calculation within participants. This test was also used to calculate the agreement between participants and radiologists as well as Volpara^{TM}. We used a

Weighted Kappa to correct for variability in the levels of disagreement between pairs of participants, given the ordinal nature of the data. Kappa was interpreted as described by Viera and Garrett[19]: < 0 (less than chance), 0.01–0.20 (slight), 0.21–0.40 (fair), 0.41–0.60 (moderate), 0.61–0.80 (substantial), 0.81–0.99 (almost perfect), and 1 (perfect) agreement.

Results

A total of 560 readings were made by participants before (pre-test) and after the-intervention (post-test). The distribution of breast density categories according to the majority report of participants relative to radiologists is shown in Figure 1.

Figure 1A and B: Distribution of cases into BI-RADS® categories according to the majority report of participants relative to radiologists.

According to the majority report of radiologists, there were 35%, 25%, 32.5%, and 7.5% of BI-RADS® 1, 2, 3, and 4 respectively. The percentage (number) of cases participants classified into different BI-RADS® categories was similar to that of the radiologists (Pre-intervention: 27.5%; 35%; 32.5%; 27.5%; 10%, and Post: 35%; 22.5%; 32.5%; 10%). However, mammogram cases classified into MBD categories differed considerably (Figure 1A and B).

Inter-reader agreement for participants before and after the intervention assessment is shown in Table 1. In the pre-intervention phase, inter-reader agreement varied from substantial (0.70; 95%CI: 0.49–0.83) to almost perfect (0.89; 95%CI: 0.79–0.94), and the overall inter-reader agreement was substantial (0.79; 95%CI: 0.70–0.87). Post-intervention analysis demonstrated substantial (0.79; 95%CI: 0.65–0.89) to almost perfect (0.92; 95%CI: 0.86–0.96) inter-reader

agreement, and an almost perfect overall agreement (0.84; 95%CI: 0.77–0.90) for all participants.

Table 1: Inter-Participant agreement (K_w) at 95% Confidence Interval (CI)

Table 2 shows the agreement between participants and radiologist as well as Volpara™. Weighted Kappa analysis demonstrated slight (0.02; 95%CI: -0.57–0.49) to fair (0.27; 95%CI: -0.37–0.61) agreement between participants and radiologists before the intervention, with a fair agreement (0.24; 95%CI: -0.46–0.61) overall. Post-intervention analysis also demonstrated slight (0.07; 95%CI: -0.77–0.51) to fair (0.32; 95%CI: -0.30–0.64) agreement, with the overall agreement between participants and radiologists being 0.31 (95%CI: -0.14–0.67). A fair agreement was observed between participants and Volpara™ in both phases: Pre-intervention (0.24; 95%CI: -0.41–0.59); Post-intervention (0.28; 95%CI: -0.34–0.61).

Table 2: Agreement (K_w) at 95%CI between participants, Radiologists and Volpara™

Discussion

Our paper explores the impact of a short self-directed, experiential learning intervention on improving reproducibility of MBD assessment. It also considers whether the intervention would improve the agreement between participants and radiologists as well as Volpara™. Overall, findings demonstrate an increase from substantial inter-reader agreement before the intervention to almost perfect agreement after the intervention. We observed a fair agreement between this cohort of participants and radiologists as well as Volpara™ before and after the intervention. Although the overall agreement between participants improved after the intervention, the improvement between a pair of participants was not consistently linear. In fact, some pairs of

participants demonstrated lesser levels of agreement after the intervention (Table 1), suggesting that not all participants benefitted from the intervention. Also, we expected that participants would demonstrate a significantly higher level of agreement with radiologists after the intervention since expert radiologists' ratings were used for the intervention. However, findings suggest that the intervention may not have significantly changed participants' perception of MBD.

Differences in subjective MBD classification has been widely reported among radiologists, with Kappa values ranging from 0.37-0.91.[10] Data produced in the current work shows variable inter-reader agreement between radiographers as well as between individual radiographers and the radiologists (Table 2). The ABR examiners from which the radiologists' data for this study was generated have also been shown to demonstrate substantial variability in MBD assessment (κ = 0.33 - 0.67),[20] albeit, with Cohen's Kappa analysis, which does not correct for the level of disagreement between pairs of assessors. Inter-reader variability in MBD assessment is attributable to observer subjectivity and differences in observer knowledge.[10] Automated tools such as Volpara™ were introduced for MBD assessment to overcome subjective variability. Our work shows poor agreement between participants and Volpara™ before and after the intervention. However, an almost perfect agreement was observed between radiologists and Volpara™. These findings are reasonable given that participants only had a short self-directed, experiential learning exercise for developing their grading skills, whereas the radiologists underwent a rigorous and lengthy training in mammography, with 24.9±8.3 mean years of specialization in mammography interpretation and a mean annual volume read of 7107±5,308.[20] In addition, Volpara™ was modelled using the MBD ratings of USA radiologists.[3] Although Volpara™ is automated and consistent in MBD assessment, the cost of installation makes its

clinical implementation difficult for some economies. Additionally, Volpara ™ does not consider all the masking effect of breast density when classifying MBD.[3] Thus, many clinical settings still rely on human subjective assessments, making interventions to improve human consistency relevant on an ongoing basis.

To ensure uniformity in MBD classification and reduce unnecessary variability in decision-making from such assessment, previous studies have recommended continuous training interventions.[11, 12] However, only one study has explored the impact of such training interventions on radiologists' reproducibility of MBD assessment.[11] The authors reported a substantial agreement (κ_w = 0.79) for radiologists trained together but did not examine readers before training, making it difficult to assess the level of improvement.[11] The current study differs from the above in that it involves diagnostic radiographers with a postgraduate mammography imaging qualification that are currently undertaking a reporting course, and assesses their inter-reader agreement before and after a self-directed learning intervention. Overall, findings show a 5% reduction in inter-reader variability after the intervention. This reduction should however not be interpreted as an improvement in their knowledge of MBD assessment. Since participants were trained using expert radiologists' rating, it would be more reasonable to assess the impact of the intervention using their level of agreement with radiologists (ABR examiners). In the current work, we observed a 7% increase in the agreement between participants and radiologists after the intervention. Our findings are similar to that of Raza et al.,[12] who reported a 7% increase in the agreement between 20 radiologists and a reference standard (expert radiologist) after a training intervention. It should be noted that the intervention was designed to get the novices working at a standard fit for practice which isn't necessarily the same as getting them working at expert level which is what they were compared against (ABR examiners).

A few factors may be responsible for the small scale of the change in agreement between participants and the radiologists. These include the limited experience of the participant cohort, their lack of familiarity with the BI-RADS® MBD rating scale, the limited intervention period, and the small sample size. The participants examined in this study were in their ninth of an eleventh month mammography reporting training programme and were being trained in the UK MBD three-point classification system (fatty, mixed, and dense). The participants had reported hundreds of cases using the UK 3-point classification scale and only received 2-3 hours training in MBD classification using the percentages described in the fourth edition BI-RADS® methodology, thus they can be classified as an inexperienced cohort. According to the memory-cueing hypothesis, training and experience influence observer performance in classification tasks.[21] Experienced observers use the memory of prior experience rather than logic to perform classification tasks; however, inexperienced observers have a conceptual overview of a task, and "tend to engage in cognitive short-circuiting responses".[21] The direct and often improper approach of inexperienced observers such as the participants in the current study may have been responsible for the low level of agreement with radiologists. It is also possible that participants were memorizing cases until their ratings matched that of ABR examiners in the intervention set rather than visually extracting the features that influenced ABR examiners' ratings. These confounding factors may have limited the value and impact of our learning intervention.

The current study is not without limitations. Firstly, we delivered the intervention over a short period, which may have affected participants' ability to master the percentages described in the fourth edition BI-RADS®. Also, we used the same test set to assess participants' pre- and post-intervention performance, albeit they had three days between pre- and post- measures to

minimize the effects of case memory. Although participants did not receive any feedback on their performance before the intervention, there is a possibility that the pre-intervention MBD assessment interfered with the independence of the post-intervention assessment. The practitioner's assessment occurred in one university in the UK, thus there were a small number of participants. Also, the number of experts evaluating the intervention images could be increased from three. However, the consensus method used by experts to agreeing on MBD reduced the impact of inter-reader variations. Further studies with larger radiographer cohort, longer periods between the base line and post intervention, as well as an enhanced self-learning experience might have a positive impact on the reproducibility of MBD assessment and further work is proposed on this basis.

Our study is novel in that it is the first to assess the impact of a self-directed experiential learning and feedback intervention in reducing inter-reader variability in radiographers for MBD classification. It is also the first to explore inter-reader agreement between qualified diagnostic radiographers who held a postgraduate qualification in mammography imaging and undergoing training in mammography reporting using a validated dataset. Findings from this work should serve as a baseline for interventions to reduce variability in MBD assessment between reporting radiographers as well as other breast readers. Improved reproducibility of MBD assessment may reduce differences in selecting women for adjunctive imaging and determining screening intervals to maximise the benefits of screening for women with dense breasts.

Conclusion

Findings demonstrate that there is substantial inter-reader agreement for MBD classification before the self-directed learning intervention. After the learning intervention inter-reader

agreement increases to almost perfect. There is fair agreement between radiographers and radiologists as well as fair agreement between radiographers and Volpara^{TM} before and after the intervention. Overall, this work demonstrates that a self-directed experiential learning and feedback intervention reduces inter-reader variability in MBD classification and that an enhanced structured training may be required to improve agreement in MBD assessment between inexperienced and expert readers. Further work is suggested for the learning intervention to improve its effectiveness.

References

1. Boyd NF, Guo H, Martin LJ, Sun L, Stone J, Fishell E, et al. Mammographic Density and the Risk and Detection of Breast Cancer. *New England Journal of Medicine.* 2007;356:227-236.

2. McCormack VA, dos Santos Silva I. Breast Density and Parenchymal Patterns as Markers of Breast Cancer Risk: A Meta-analysis. *Cancer Epidemiology Biomarkers & Prevention.* 2006;15:1159-1169.

3. Ekpo EU, Hogg P, Highnam R, McEntee MF. Breast composition: Measurement and clinical use. *Radiography.* 2015;21:324-333.

4. Yaghjyan L, Mahoney MC, Succop P, Wones R, Buckholz J, Pinney SM. Relationship between breast cancer risk factors and mammographic breast density in the Fernald Community Cohort. *Br. J. Cancer.* 2012;106:996-1003.

5. Tice JA, Cummings SR, Ziv E, Kerlikowske K. Mammographic breast density and the Gail model for breast cancer risk prediction in a screening population. *Breast Cancer Res Treat.* 2005;94:115-122.

6. Warwick J, Birke H, Stone J, Warren RM, Pinney E, Brentnall AR, et al. Mammographic breast density refines Tyrer-Cuzick estimates of breast cancer risk in high-risk women: findings from the placebo arm of the International Breast Cancer Intervention Study I. *Breast Cancer Res.* 2014;16:451.

7. Ekpo EU, Brennan PC, Mello-Thoms C, McEntee MF. Relationship Between Breast Density and Selective Estrogen-Receptor Modulators, Aromatase Inhibitors, Physical Activity, and Diet: A Systematic Review. *Integrative cancer therapies.* 2016;15:127-144.

8. Eriksson. L, Czene. K, Rosenberg. LU, Törnberg. S, Humphrey. K, Hall. P. Mammographic density and survival in interval breast cancers. *Breast Cancer Research.* 2013;15.

9. Haas JS, Kaplan CP. The Divide Between Breast Density Notification Laws and Evidence-Based Guidelines for Breast Cancer Screening: Legislating Practice. *JAMA internal medicine.* 2015;175:1439-1440.

10. Ekpo EU, McEntee MF. Measurement of breast density with digital breast tomosynthesis--a systematic review. *The British journal of radiology.* 2014;87:20140460.

11. Ekpo EU, Ujong UP, Mello-Thoms C, McEntee MF. Assessment of Interradiologist Agreement Regarding Mammographic Breast Density Classification Using the Fifth Edition of the BI-RADS Atlas. *AJR. American journal of roentgenology.* 2016;206:1119-1123.

12. Raza S, Mackesy MM, Winkler NS, Hurwitz S, Birdwell RL. Effect of Training on Qualitative Mammographic Density Assessment. *Journal of the American College of Radiology : JACR.* 2016;13:310-315.

13. Wade RC YD. Portfolios: A tool for reflective thinking in teacher education? *Teaching Teacher Educ.* 1996;12:63-79.

14. Aukes LC, Geertsma J, Cohen-Schotanus J, Zwierstra RP, Slaets JPJ. The Effect of Enhanced Experiential Learning on the Personal Reflection of Undergraduate Medical Students. *Medical Education Online.* 2008;13:15.

15. Ertmer PA, Newby TJ. The expert learner: Strategic, self-regulated, and reflective. *Instructional Science.* 1996;24:1-24.

16. Schonrock-Adema J, Heijne-Penninga M, van Duijn MA, Geertsma J, Cohen-Schotanus J. Assessment of professional behaviour in undergraduate medical education: peer assessment enhances performance. *Medical education.* 2007;41:836-842.

17. Culpan AM. Radiographer involvement in mammography image interpretation: A survey of United Kingdom practice. *Radiography.*22:306-312.

18. Torres-Mejía G, Smith RA, Carranza-Flores MdlL, Bogart A, Martínez-Matsushita L, Miglioretti DL, et al. Radiographers supporting radiologists in the interpretation of screening mammography: a viable strategy to meet the shortage in the number of radiologists. *BMC Cancer.* 2015;15:410.

19. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. *Family medicine.* 2005;37:360-363.

20. Damases CN, Brennan PC, Mello-Thoms C, McEntee MF. Mammographic Breast Density Assessment Using Automated Volumetric Software and Breast Imaging Reporting and Data System (BIRADS) Categorization by Expert Radiologists. *Academic radiology.* 2016;23:70-77.

21. Cox JR, Griggs RA. The effects of experience on performance in Wason's selection task. *Memory & Cognition.* 1982;10:496-502.

Figure Legend

Figure 1: Distribution of cases into BI-RADS® categories according to the majority report of participants relative to radiologists. Black bars represent the majority report of ABR examiners and the grey bars represent that of radiographers for each case in the test set.