Research Open

Interventional Medicine and Clinical Imaging

Volume 1 Issue 1

Research Article

Resident and Faculty Concordance in Screening Mammography: Impact of Experience and Opportunities for Focused Instruction

Anjali L. Saripalli[1], Katherine A. Klein[2]*, Mark A. Helvie[1], Janet E. Bailey[1], Colleen Neal[1], Stephanie K. Patterson[1], Renee W. Pinsky[1] and Katherine E. Maturen[1]

[1]Department of Radiology, Michigan Medicine-University of Michigan, 2910A TC SPC 5326, 1500 E. Medical Center Drive, Ann Arbor, MI, 48109, USA

[2]Department of Radiology, Michigan Medicine-University of Michigan, 2910A TC SPC 5326, 1500 E. Medical Center Drive, Ann Arbor, MI, 48109, USA

***Corresponding author:** Katherine A. Klein, Department of Radiology, Michigan Medicine-University of Michigan, 2910A TC SPC 5326, 1500 E. Medical Center Drive, Ann Arbor, MI, 48109, USA; **Email:** kleink@med.umich.edu

Received: December 07, 2018; **Accepted:** December 16, 2018; **Published:** December 18, 2018;

Abstract

Purpose: To evaluate the frequency of and reasons for patient callback from offline screening mammography, comparing residents and breast imaging faculty.

Methods: Residents and MQSA-approved fellowship-trained breast imaging faculty independently recorded prospective interpretations of a subset of bilateral clinical screening mammograms performed over a 1-year period at our NCI-designated cancer site utilizing Computer-Assisted Diagnosis (CAD). BI-RADS 1, 2, or 0 were allowed at screen interpretation. IRB-approved retrospective review compared callback performance in both groups. Descriptive statistics and multivariate logistic regression were performed.

Results: 1317 consecutive bilateral screening mammograms were reviewed. Residents recommended callback for 123/1317 (9.3%) and faculty for 110/1317 (8.4%) women (p<.0001). Overall agreement was moderate (k=0.50) with lower agreement between faculty and novices (experience < 4 weeks) (k=0.39) than between faculty and senior residents (experience > 8 weeks) (k=0.63). Agreement varied with findings: calcifications (k=0.66), mass (k=0.52), focal asymmetry (k=0.45), asymmetry (k=0.33). In multivariate regression, all four finding types were predictors of discordance: calcifications (OR 10.4, 95% CI 3.4, 33.1, p<.0001); mass (OR 19.2, 95% CI 7.7, 48.0, p<.0001); focal asymmetry (OR 21.3, 95% CI 9.9, 45.7, p<.0001); asymmetry (OR 40.1, 95% CI 21.4, 75.2, p<.0001). Odds of discordance declined by 6% with each week of resident experience (OR 0.94, 95% CI 0.89, 0.99, p=.02). Breast density was not a significant predictor.

Conclusions: Resident and faculty callback agreement was moderate but improved with resident experience. Novices often detected calcifications and masses but missed focal asymmetry and asymmetry, suggesting educational efforts should focus on the perception of asymmetry.

Introduction

Breast cancer has the second highest mortality rate of all cancers in women, and mammography is the only known screening method shown to decrease disease-related mortality [1]. Robust diagnostic performance of screening mammography is essential to this public health impact, with a delicate balance between detecting clinically significant cancers and avoiding excessive callback rates. This level of accuracy is the intended result of specialty training and years of experience in breast imaging, but the first phase is residency training [2, 3].

To meet the requirements of the Mammography Quality Standards Act (MQSA) for training in breast imaging, radiology residents spend at least 12 weeks of their 4 year training in breast imaging clinical rotations [4]. Resident evaluations are based on faculty observation of interpretative skills and procedures, patient interactions, and dictated reports. This style of individualized instruction has the potential to provide residents with personalized training. However, given the time constraints often present at busy academic centers, there is a further need for objective metrics and data that can be used to assess the performance and tailor the education of trainees in breast imaging.

As residency training is integral to mammography expertise, many previous efforts have focused on improving the training process. Previous efforts have addressed the need for varied difficulty of cases based on self- and expert-assessments to maximize the effect of training on resident performance [5]. Mathematical models have been developed in an effort to address the need for objective assessment metrics [6, 7], and some efforts have been made to identify image features predictive of error to improve the clinical utility of such models [8].

Outside of breast imaging, concordance of resident and faculty interpretation is high [9, 10]. That is not the case in breast imaging. The goal of the current study is to evaluate the frequency and morphologic reason for trainee callbacks from screening mammography and to compare them to faculty breast imager callbacks. We hypothesize that the callback rates of radiology residents will be within the national benchmarks of 8–12% but higher than those of experienced breast imaging faculty.

Materials and Methods

All cases interpreted were 2D digital four view screening mammograms obtained on GE Senographe Essential Mammography equipment (Buc, France) at one of six screening locations within one academic health system. Residents and faculty had individual workstations to view the digital studies with hard copy images available for review as desired.

Anonymized screening mammography data sheets, including resident, and faculty interpretations, were routinely recorded for Quality Assessment (QA) and educational purposes from July 1, 2014, to June 30, 2015. All the radiology residents who rotated in breast imaging took part in this process. It has been shown that trainee interpretation of screening mammography influences faculty interpretation [2]. Thus, we asked residents and faculty to fill out an initial written assessment form independently, stating whether they would recall the mammography patient for additional screening or interpret the mammogram as negative. Faculty interpretation was the reference standard for purposes of this study. Subsequent Institutional Review Board (IRBMED) approval for retrospective reviews of the data waived the need for patient consent. Data included resident weeks of training, resident observations (calcifications, mass, focal asymmetry, asymmetry), location, recommendations for a callback for additional diagnostic imaging as well as faculty observations, location, recommendation, and assessment of breast density. All eleven faculties in the breast imaging section, with nine to thirty years of experience after fellowship, were included.

The hard copy data were subsequently entered into an electronic spreadsheet by a medical student blinded to clinical outcomes (Microsoft Excel, Redmond, WA). Resident interpretation was considered concordant with faculty interpretation when the decision and reason for callback matched that of the faculty, for one breast in per breast analysis or both breasts for per patient analysis. Descriptive statistics were performed to identify data trends and distribution. Continuous variables were evaluated with means and compared using t-tests or non-parametric tests where appropriate, while categorical variables were expressed as counts or percentages and compared using chi-square tests and measures of agreement. Kappa agreement was considered slight if <.20, fair if 0.21–0.40, moderate if 0.41–0.60, substantial if 0.61–0.80, and almost perfect if 0.81–0.99. Logistic regression analysis was performed to evaluate predictors of resident-faculty discordance. A stepwise forward selection algorithm was used to select covariates for multivariate logistic regression. All statistical procedures considered p<.05 as the standard for statistical significance and were performed using SAS 9.4 (SAS Institute, Cary, NC).

Results

Data sheets were reviewed for 1,345 consecutive bilateral screening mammograms; 28 of these were excluded from further analysis because the data sheets were incomplete (n=27), or the patient had clinical symptoms that would warrant a diagnostic exam regardless of screening mammographic findings (n=1), leaving 1,317 cases. Residents recommended that 123/1,317 (9.34%) women be called back for additional imaging, while faculty recommended callbacks for 110/1,317 (8.35%) women (p<.0001). Resident and faculty callback recommendations at the per-patient level were concordant in 1208/1,317 (91.72%) cases. Residents and faculty agreed on 62 callbacks, while residents would have called back 61 women who were not called back by faculty, and faculty called back 48 women who would not have been called back by residents. Among the 62 cases of apparently concordant callbacks, the sidedness of the resident and faculty's reasons for callback differed in 5/62 (8.07%) cases. Therefore, the true proportion of concordant interpretations on the per-patient level was 91.34%, and the remaining analysis was performed on a per-breast basis with a total sample size of 2,634.

Regarding each breast as an individual observation, the residents recommended callback in 139/2634 (5.28%) cases and the faculty in 123/2634 (4.67%) cases (p<.0001). Overall agreement between residents and faculty was moderate (k=0.50, p<.0001). Recommendations were negative concordant (no call back) in 2441/2634 (92.67%) cases, positive concordant (both call back) in 69/2634 (2.63%), resident positive/faculty negative in 70/2634 (2.66%) and resident negative/faculty positive in 54/2634 (2.05%). Types and locations of findings prompting callbacks are illustrated in Figures 1 and 2. Resident and faculty agreement were highest for calcifications (k=0.66) and lowest for asymmetry (k=0.33), presented in table 1. Agreement for location was moderate (k=0.45).

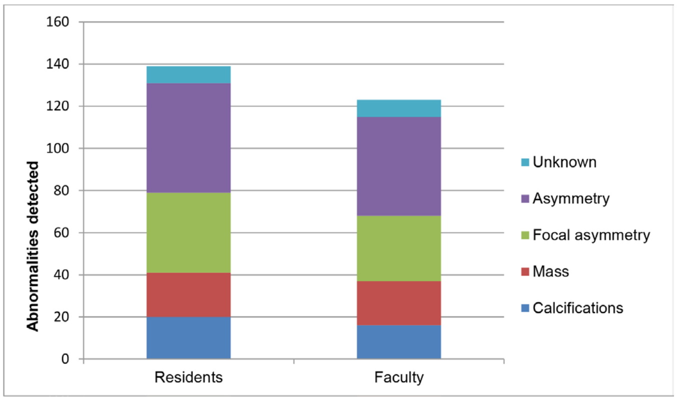

Figure 1. M. ammographic findings prompting recommendation for callbacks among residents and faculty, on a per breast basis.

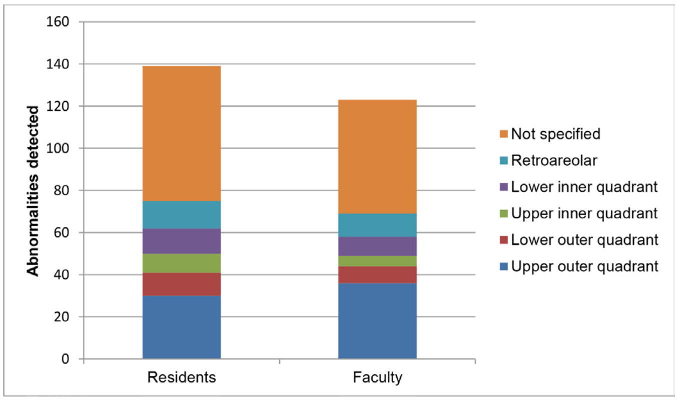

Figure 2. Location of findings prompting recommendation for callbacks among residents and faculty, on a per breast basis

Table 1. Agreement between residents and faculty on type and location of findings prompting recommendation for callback from screening mammography, on a per breast basis. P-values < .05 indicate the presence of a non-zero correlation between faculty and trainee interpretations of each feature.

	Cohen's kappa	*p* value
Calcifications	0.66	<.0001
Mass	0.52	<.0001
Focal asymmetry	0.45	<.0001
Asymmetry	0.33	<.0001
Location	0.45	<.0001

Breast composition was classified by faculty in 2035 cases, by ACR BI-RADS v.5 (ACR 2013). 322/2035 (15.82%) were almost entirely fatty (A); 961/2035 (47.22%) had scattered areas of fibro glandular density (B); 690/2035 (33.90%) were heterogeneously dense (C); and 62/2035 (3.06%) were extremely dense (D).

1054/2634 (40.02%) of cases were read by a first-year radiology resident, 542/2634 (20.58%) by a second-year resident, 30/2634 (1.14%) by a third-year resident, and 1008/2634 (38.27%) by a fourth-year resident. Residents had 0–15 weeks (mean 6.11 ± 3.96 weeks) of prior experience in breast imaging.

Univariate logistic regression analysis was performed to evaluate whether any of the following features was a significant predictor of resident-faculty discordance: any of the four major types of findings (as judged by faculty), the presence of moderately (Classifications C+D vs. A+B) or extremely (Classification D vs. A+B+C) dense breasts, or the duration of the resident's breast imaging experience. These results are presented in table 2.

Table 2. Parameter estimates from univariate logistic regression predicting resident-faculty callback discordance.

	Outcome: Discordance		
	Odds ratio	95% CI	*p* value
Calcifications	6.94	2.21, 21.83	<.001
Mass	13.24	5.38, 32.59	<.0001
Focal asymmetry	14.05	6.66, 29.65	<.0001
Asymmetry	28.55	15.56, 52.40	<.0001
Moderately dense breasts	1.45	1.00, 2.11	0.05
Extremely dense breasts	0.67	0.16, 2.77	0.58
Resident experience (unit = 1 week)	0.96	0.91, 1.00	0.09

Multivariate logistic regression of all factors was performed using stepwise forward selection, and all four types of findings, as well as resident experience, were retained as significant predictors. The purpose of multivariate regression is to control for other factors that may alter the odds ratio estimates of each parameter. Parameter estimates are presented in Table 3.

Discussion

Our retrospective analysis of resident and faculty callbacks in 1345 screening mammograms demonstrated moderate agreement (k=0.50) between residents and faculty. Residents recommended callback more frequently than faculty (9.34% vs. 8.35% of women, p<.0001). Radiology residents are aware of the national benchmark for screening breast mammography callbacks, which could explain the low rate. Agreement improved with resident experience so that the odds of discordance dropped by 6% for every week of resident experience in multivariate analysis. All four major types of findings prompting callbacks were associated with discordance. The Kappa agreement was highest for the presence of calcifications (k=0.66) and lowest for asymmetry (k=0.33) with the higher concordance for the presence of calcifications possibly related to presence of coronary artery disease. Likewise, the odds ratios for discordance ranged from 10.39 (95% CI 3.27, 33.08, p<.0001) for calcifications to 40.10 (95% CI 21.38, 75.21, p<.0001) for asymmetry. Breast density was not a significant predictor of discordance.

Table 3. Parameter estimates from multivariate logistic regression predicting resident-faculty discordance.

	Outcome: Discordance		
	Odds ratio	95% CI	*p* value
Calcifications	10.39	3.27, 33.08	<.0001
Mass	19.23	7.71, 47.96	<.0001
Focal asymmetry	21.31	9.92, 45.74	<.0001
Asymmetry	40.10	21.38, 75.21	<.0001
Resident weeks of experience (unit = 1week)	0.94	0.89, 0.99	0.02
C statistic	0.70		

Benchmark Comparison

In the highly regulated and monitored world of screening mammography, recall rate is a performance metric that has been included in most accreditation guidelines. It is easy to obtain and has been used to assess institutional and personal professional quality. In our study, recall rate is defined as the number of screening studies with a final recommendation of BI-RADS 0 (Incomplete: needs additional imaging evaluation) out of the entire screening pool.

The 2017 update to the Breast Cancer Surveillance Consortium (BCSC) benchmarks for screening mammography is essential because it reflects modern technology and practice methods. In this study, only 59% of the radiologists studied fell within the national benchmark recall range of 5–12% with a trend towards higher recall rates [11]. The National Mammography Database (NMD) is a mammography data registry also providing performance metrics for clinical practice [12] that reported a mean recall rate of 10% from the NMD with a range of 8–11.4% based on practice location and type (using comparable BI-RADS 4 recall inclusion definition). The mean recall rate in an academic setting was 9.8%.

Our data show that the recall rates for the faculty (8.35%) and residents (9.34%) both fall within the benchmark ranges by national and academic center standards. As a QA measure, this is important and timely as this is a potential metric proposed by the Physician

Quality Reporting System (PQRS) by the Centers for Medicare and Medicaid Services to determine payment for services [12]. [13] performed a reader study to assess the accuracy of interpretation of screening mammograms, concluding that diagnostic volume was not the only contributor to performance. Instead, they posited a multifactorial process that they could not yet fully define. Thus, the difference in recall rate between faculty and residents in the current study is unlikely to arise from differences in interpretation volume alone.

Discordancy Rates

In the literature, interest in the concordance of radiology resident image interpretation compared to faculty interpretation has focused on residents' on-call interpretations.

Discordance has been shown to vary depending on the complexity of imaging. MRI cases, followed by CT, are the most common sources of discordant resident interpretations. Next, plain radiographs are the third most likely image type to be associated with discordance, followed by ultrasound, a modality where residents may be helped by experienced technologists [10, 14].

Discordance on call has been shown to decrease as residents progress in their training, presumably because resident knowledge and skill improve with clinical experience and didactics [14]. However, it has been shown that subspecialist breast imagers detect more cancers (and more early-stage cancers) and have lower recall rates than general radiologists [15]. Towards the end of their training, radiology residents are largely comparable to novice general radiologists. In agreement with Lewis et al, we found that residents with more breast imaging experience were more concordant with breast imaging subspecialty faculty [7] It is likely that the subtler finding of mammographic asymmetry, which was associated with the largest odds of discordance, requires more experience for reliable detection than a discrete finding like calcifications.

Limitations

This retrospective study is subject to several limitations. First, the data collection method does not allow for the identification of the resident or faculty, so it is not possible to control for the intrinsic correlation between multiple readings by the same person. Instead, each mammographic interpretation is treated as an independent observation, which could impact both confidence intervals and overall statistical inference. Otherwise stated, a specific radiologist's tendency to overcall or under call may be a more powerful predictor than his or her level of training or the patient's breast density, but we are unable to test for this. Second, patient age was not included on the data sheets but could have been a factor affecting clinical interpretations either consciously or unconsciously. Third, the experience level of the faculty was not noted on the data collection sheets, but given that all of the faculty involved were at least nine years out of training, this is considered to be a minor issue. Finally, the anonymized data collection method does not enable linkage of the screening mammogram to the results of any subsequent diagnostic workup, so the clinical significance of any resident-faculty discordance remains unknown.

Conclusion

We compared frequency and rationale for callbacks from offline screening mammography between residents and breast imaging faculty and found that while resident and faculty callback agreement was only moderate, it improved with resident experience. While novices often detected calcifications and masses, concordance was low for the more subtle findings of asymmetry, suggesting educational efforts should increase emphasis on the perception of asymmetry.

References

1. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention (CDC) and the National Cancer Institute (NCI) (2015) 1999–2013 Cancer Incidence and Mortality Data. Available from: https://nccd.cdc.gov/uscs/.
2. Hawley JR, Taylor CR, Cubbison AM, Erdal BS, Yildiz VO, et al. (2016) Influences of Radiology Trainees on Screening Mammography Interpretation. *J Am Coll Radiol* 13: 554–561. [crossref]
3. Poot JD, Chetlen AL (2016) A Simulation Screening Mammography Module Created for Instruction and Assessment: Radiology Residents vs National Benchmarks. *Acad Radiol.* 23: 1454–1462.
4. Sickles EA, Philpotts LE, Parkinson BT, Monticciolo DL, Lvoff NM, Ikeda DM, et al. (2006) American College Of Radiology/Society of Breast Imaging curriculum for resident and fellow education in breast imaging. *J Am Coll Radiol.* 3: 879–884.
5. Grimm LJ, Kuzmiak CM, Ghate SV, Yoon SC, Mazurowski MA (2014) Radiology resident mammography training: interpretation difficulty and error-making patterns. *Acad Radiol.* 21: 888–892.
6. Wang M, Wang M, Grimm LJ, Mazurowski MA (2016) A computer vision-based algorithm to predict false positive errors in radiology trainees when interpreting digital breast tomosynthesis cases. *Expert Systems with Applications* 64: 490–499.
7. Lewis PJ1, Rooney TB2, Frazee TE2, Poplack SP3 (2018) Assessing Resident Performance in Screening Mammography: Development of a Quantitative Algorithm. *Acad Radiol* 25: 659–664. [crossref]
8. Grimm LJ, Ghate SV, Yoon SC, Kuzmiak CM, Kim C, et al. (2014) Predicting error in detecting mammographic masses among radiology trainees using statistical models based on BI-RADS features. *Med Phys* 41: 031909. [crossref]
9. Xiong L, Trout AT, Bailey JE, Brown RKJ, Kelly AM (2011) Comparison of Discrepancy Rates in Resident and Faculty Interpretations of On-Call PE CT and V/Q Scans: Is One Study More Reliable During Off Hours? *Journal of the American College of Radiology* 8: 415–421.
10. Ruma J, Klein KA, Chong S, Wesolowski J, Kazerooni EA, et al (2011) Cross-sectional examination interpretation discrepancies between on-call diagnostic radiology residents and subspecialty faculty radiologists: analysis by imaging modality and subspecialty. *J Am Coll Radiol* 8: 409–414.
11. Lehman CD, Arao RF, Sprague BL, Lee JM, Buist DS, Kerlikowske K, et al (2017) National Performance Benchmarks for Modern Screening Digital Mammography: Update from the Breast Cancer Surveillance Consortium. *Radiology* 283: 49–58.
12. Lee CS, Bhargavan-Chatfield M, Burnside ES, Nagy P, Sickles EA (2016) the National Mammography Database: Preliminary Data. *AJR Am J Roentgenol* 206: 883–890. [crossref]
13. Beam CA, Conant EF, Sickles EA (2003) Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. *J Natl Cancer Inst* 95: 282–290. [crossref]
14. Weinberg BD, Richter MD, Champine JG, Morriss MC, Browning T (2015) Radiology resident preliminary reporting in an independent call environment: multiyear assessment of volume, timeliness, and accuracy. *J Am Coll Radiol* 12: 95–100.
15. Sickles EA, Wolverton DE, Dee KE (2002) Performance parameters for screening and diagnostic mammography: specialist and general radiologists. *Radiology* 224: 861–869. [crossref]

Citation:
Saripalli AL, Klein KA, Helvie MA, Bailey JE, Neal C, Patterson SK, Pinsky RW, Maturen KE(2018) Resident and Faculty Concordance in Screening Mammography: Impact of Experience and Opportunities for Focused Instruction. *Interv Med Clin Imaging* Volume 1(1): 1–5.

Appendix

Location_____________ Radiology Residency Year_______________ No. week's experience__________

Screener	**Resident negative**	**Resident Callback** 1=Ca++ 2=Mass 3=Focal Asymmetry 4=Asymmetry 5=Diagnostic	**Location** 1=UOQ 2=LOQ 3=UIQ 4=LIQ 5=RetroA 6=DNK	**Faculty negative**	**Faculty Callback** 1=Ca++ 2=Mass 3=Focal Asymmetry 4=Asymmetry 5=Diagnostic	**Location** 1=UOQ 2=LOQ 3=UIQ 4=LIQ 5=RetroA 6=DNK	IF BOTH CALLBACK WAS IT FOR THE SAME REASON?	**Density** 1=Fatty 2=Scattered 3=Hetero- Dense 4=Extreme Dense
1 R								
L								
2 R								
L								
3 R								
L								
4 R								
L								
5 R								
L								

Comments____

LOCATION	NUMBER
Upper outer Quadrant	1
Upper Inner Quadrant	2
Lower Outer Quadrant	3
Lower Inner Quadrant	4
Retroareolar	5
***DNK**	6
*Do Not Know because only seen on one view	

CALL BACK REASON	NUMBER
Ca++	1
Mass	2
Focal Asymmetry	3
Asymmetry	4
Architectural Distortion	5
Diagnostic Reason	6