Recall Laterality and Bilaterality: Possible New Screening Mammography Quality Metrics

Authors: *Samson Munn,[1] Virginia Huynh Kim,[1,2] Joanna Huijia Chen,[1,3] Sean Maldonado Ramirez,[1,4] Michelle Kim,[1] Paul Koscheski,[1] Babak N. Kalantari,[1] Gregory Eckel,[1] Albert Lee[1]

1. Department of Radiology, Harbor-UCLA Medical Center, Torrance, California, USA
2. Department of Radiology, Kaiser Permanente Medical Center, San Leandro, California, USA
3. Bay Imaging Consultants Medical Group, Walnut Creek, California, USA
4. Department of Radiology, Holy Cross Health, Fort Lauderdale, Florida, USA

*Correspondence to samsonmunn@pm.me

Disclosure: Munn has a patent issued for a method and system for evaluating the performance of a reader of screening mammograms (patent 11,037,086). The authors declare no other conflicts of interest.

Received: 26.02.2024

Accepted: 13.06.2024

Keywords: Audit, breast cancer, Breast Imaging-Reporting and Data System (BI-RADS), mammography, Mammography Quality Standards Act (MQSA), screening quality.

Citation: Oncol AMJ. 2024;1[1]:73-80. https://doi.org/10.33590/oncolamj/MZKD5370.

Abstract

Purpose: Current screening mammography quality metrics are important and helpful, but do not address all quality concerns. An individual screening mammography reader may be systematically insensitive to findings present in the breast of one side, laterality bias, evidenced by left versus right difference in advised immediate recalls. Current metrics are not designed to detect laterality bias. Whether a reader exhibits laterality bias, or what an appropriate ratio/range of bilateral versus unilateral recalls should be, have never been discussed in medical literature.

Methods: As a trainee quality project, five attending ('consultant' in Europe) radiologists' screening mammography reports over 2 years at an academically affiliated, public hospital were tallied with regard to laterality of recommended recall, and with respect to unilateral versus bilateral recalls advised. The chi-square (χ^2) statistic was applied to reports advising unilateral recall.

Findings: No group laterality bias was discovered. One radiologist (the most experienced) evidenced a consistent laterality bias over 2 years ($p=0.07$) against left-breast findings. Of reports recommending recall, the radiologists' single-year range for recall regarding both breasts was 10.2–23.3%; for both years combined, the individual radiologists ranged from 13.6–17.9%. The group, 2-year mean recommending bilateral recall was 16.5%.

Conclusion: A radiologist may exhibit laterality bias, favoring detection of findings in one

breast over the other, a concern never before considered. Audit to discern such bias leads simultaneously to assessment of bilateral recall bias. Possible causes of these biases are discussed, and research regarding them as possible quality metrics is encouraged.

Key Points

1. Laterality bias: There is no known, prior medical literature regarding whether a reader of screening mammograms might render interpretations with bias toward detection of findings on one side (left versus right). In a simple audit of mammogram reports made by five radiologists over 2 years, a strong, unexpected likelihood that one of them exhibited laterality bias was discovered.

2. Bilaterality bias: In screening mammogram reports that recommend recall of patients for further assessment, a tendency for those recommendations to be of both breasts rather than simply one may be termed bilaterality bias. Laterality and bilaterality biases may coexist.

3. Screening mammography quality metrics: How far a reader's unilateral recall recommendations may appropriately diverge from 50–50 (reflecting laterality bias), and how small the fraction of recall recommendations for further assessment of both breasts should be (reflecting bilaterality bias), have the potential to become meaningful, practical, and easily audited new quality metrics in screening mammography.

INTRODUCTION

Quality is important in breast imaging, and there exist a number of excellent screening mammography quality metrics to audit readers and mammography programs.[1,2] Even commercially available software programs used in mammography reporting include the capability to apply metrics.[3] Although highly useful, current metrics do not account for all performance variations, nor does any single metric assess the entire screening episode.[4] Thus, there may be potential benefit in additional quality metrics.

Often, a screening mammogram report includes a recommendation for the patient to return for additional imaging. These are termed 'recall' or 'call-back' examinations. One commonly utilized metric of quality is the recall rate: the proportion of screening mammogram reports that are positive. The recall rate in the Breast Imaging Reporting and Data System (BI-RADS)[5] is calculated as the total of BI-RADS 0, 3, 4 and 5 reports divided by the number of screening exams reported.[6,7] However, as a practical matter, screening mammograms are essentially never given BI-RADS categorizations of 3, 4 or 5; so, that recall rate effectively becomes reflected by BI-RADS 0 categorizations, which refer to recommendations to obtain additional imaging or prior exams with which to compare soon.

Two potential metrics had years earlier been conceived by one of the authors: A) whether a screening mammogram reader might be biased in terms of laterality, and B) among recalls advised, whether the unilateral versus bilateral proportion is appropriate (herein termed bilaterality bias). There is no English-language medical literature regarding what portion of screening recalls should be bilateral, nor how much unilateral recalls may diverge with quality from 50–50, left–right. If such biases were to exist, their early detection could be beneficial toward identifying the underlying cause and its remediation, as quality improvement measures.

Therefore, as a trainee quality preliminary project, radiology residents simply tallied these in the authors' department, as reflected in BI-RADS 0 reports. The aim of the tally was to discover if such bias existed in the authors' department of radiology.

METHODS

A simple, observational, retrospective tally was done of bilateral, screening mammogram BI-RADS 0 interpretations made by five attending radiologists at the authors' academically affiliated, public (county), general hospital, from September 1st 2015–August 31st 2016, and from September 1st 2016–August 31st 2017: total reports, number advising recall, unilateral recalls advised for each breast, and bilateral recalls advised. The chi-square (χ^2) statistic was applied to unilateral recall reports, regarding left versus right breasts. The χ^2 statistic is appropriate to assess whether the difference from exactly 50% is one which may reasonably be expected simply randomly; in other words, how often such difference would likely be due to genuine bias (of some sort) versus simple, statistical randomness. All mammograms of the tallied reports had been screening, bilateral, and digital, and had included two routine complementary views per breast (mediolateral oblique and craniocaudal); none had included tomosynthesis. Since report validity was not being audited, mammogram images were not accessed, viewed nor correlated with the reports. The hospital is large (550 beds) and has its own, fully accredited, radiologist training program, and also trains one 'fellow' (already a board-certified radiologist) per year in the subspecialty of breast imaging, including mammography. There were five consultant, specialist radiologists who trained the residents and fellow in breast imaging. One radiologist had over 40 years of experience in breast imaging, including some years heading the Breast Imaging Division of the Department of Radiology; since that experience began before breast imaging fellowships were common, that radiologist was not fellowship-trained. The remaining four radiologists were all fellowship-trained in breast imaging, one 6 months, two 12 months, and one 17 months, and their post-fellowship breast imaging experience varied from 2–9 years prior to the audit period. Each radiologist met all the requirements of the Mammography Quality Standards Act[8] before, during, and immediately after the audit period, including having "interpreted or multi-read at least 960 mammographic examinations" (screening plus diagnostic) each 24 months. Finally, mammography at the hospital was accredited by the American College of Radiology (ACR), including with regard to the five consultant, specialist mammography radiologists.

Although not required, informed consent was obtained from the four radiologists alive when this work was done, and from the next-of-kin of the one radiologist who had since died. Informed consent and ethical approval at the authors' hospital were not required because this work was a simple retrospective tally exclusively of report categorizations, was a means for radiologist trainees to satisfy a training program requirement in quality assessment and/or improvement, did not entail an intervention nor clinical trial, did not entail patients themselves nor access to their mammogram images, did not assess the clinical accuracy of mammogram reports, did not record patient-identifying data, and did not meet the criteria of the official USA governmental definition of "human studies research",[9] at a minimum because it was not intended to be "generalizable" (i.e., to be disseminated) when it was conducted. Since it was, by definition, not human studies research, human 'subjects' were therefore not involved. For such simple radiology residency trainee projects that do not involve patients nor their images, and in which data are not recorded in identifiable fashion, institutional review board submission and ethical approval are generally not required in USA radiology training programs.

RESULTS

Over the 2 years, a total of 4,771 screening mammogram reports were audited. The data regarding reports by the radiologists as a group are in Table 1. Group sidedness disparity was not significant: the lowest group p-value Years 1, 2, and 1+2 was >0.40. Of interpretations recommending recall, the

Table 1: Radiologist group, screening mammogram report data.

	Total Number of Screening Mammogram Reports	Screening reports advising recall (recall rate)	Screening Reports Advising Recall (Recall Rate) (%)	Of Screening Reports Advising Recall, Those for One Breast (%)	Of Screening Reports Advising Unilateral Recall, Those for the Right Breast (%)	Of Screening Reports Advising Unilateral Recall, Those for the Left Breast (%)	Mean (Median) Number of Screening Mammogram Reports Per Radiologist, Per Year
Year 1	2,665	556 (20.86)	99 (17.81)	457 (82.19)	235 (51.42)	222 (48.58)	533 (545)
					p=0.54		
Year 2	2,106	439 (20.85)	65 (14.81)	374 (85.19)	192 (51.34)	182 (48.66)	421 (433)
					p = 0.61		
Years 1+2	4,771	995 (20.86)	164 (16.48)	831 (83.52)	427 (51.38)	404 (48.62)	477 (448)
					p=0.42		

2-year group mean recommending bilateral recall was 16.5%. The results regarding the individual radiologists' reports are in Table 2. Regarding four of the radiologists, the lowest (most significant) single-year p-value was 0.42; it was likewise high (i.e., insignificant) for the combined 2 years at 0.60. In contrast, the unilateral recall reports of the remaining radiologist (#3 in Table 2) disproportionately concerned the right breast (p=0.07), suggesting unilateral bias. This radiologist was the most experienced, interpreted the largest number of exams each year, and had the lowest overall recall rate. Of reports advising recall, the radiologists' individual, bilateral recall ranges were 13.2–23.3% (13.3–23.3% excluding the one radiologist with apparent laterality bias), 10.2–22.5%, and 13.6–17.9%, for Years 1, 2, and 1+2, respectively. The whole-group, 2-year mean recommending bilateral recall was 16.5% (16.1% excluding the one radiologist with apparent laterality bias).

DISCUSSION

The exclusive intent of this tally was practical and simple: to discover by solely counting screening mammography reports if there might exist laterality bias in screening mammogram reports in the authors' radiology department. Thus, patient images, patient outcomes, report validity, finding-type, and radiologist confidence levels were not considered. This audit's nature was preliminary: to gather data; there was no intervention intended nor employed.

For any group of exams interpreted, over any substantial period of time, it is unlikely that left versus right findings will lead to recall precisely 50.00% equally, left versus right; there will nearly always be a small, arithmetic side-discrepancy. Overall, the laterality data did not reveal a concern until drilling down to individual readers.

A consistent bias against one side in recall was evidenced by just one of the five

Table 2: Individual radiologist, screening mammogram report data.

Radiologist	Audit Year(s)	Total Number of Reports	Reports Advising Recall (Recall Rate) (%)	Reports Advising Bilateral Recall (%)	Reports Advising Unilateral Recall (%)	Reports Advising Unilateral Recall, Right Breast (%)	Reports Advising Unilateral Recall, Left Breast (%)
1	1	458	92 (20.09)	17 (18.48)	75 (81.52)	41 (54.67)	34 (45.33)
						p=0.42	
	2	272	59 (21.69)	6 (10.17)	53 (89.83)	26 (49.06)	27 (50.94)
						p=0.89	
	1+2	730	151 (20.68)	23 (15.23)	128 (84.77)	67 (52.34)	61 (47.66)
						p=0.60	
2	1	545	105 (19.27)	14 (13.33)	91 (86.67)	45 (49.45)	46 (50.55)
						p=0.92	
	2	389	94 (24.16)	13 (13.83)	81 (86.17)	43 (53.09)	38 (46.91)
						p=0.58	
	1+2	934	199 (21.31)	27 (13.57)	172 (86.43)	88 (51.16)	84 (48.84)
						p=0.76	
3	1	685	106 (15.47)	14 (13.21)	92 (86.79)	52 (56.52)	40 (43.48)
						p=0.21	
	2	574	89 (15.51)	20 (22.47)	69 (77.53)	40 (57.97)	29 (42.03)
						p=0.19	
	1+2	1,259	195 (15.49)	34 (17.44)	161 (82.56)	92 (57.14)	69 (42.86)
						p=0.07	
4	1	587	150 (25.55)	30 (20.00)	120 (80.00)	61 (50.83)	59 (49.17)
						p=0.86	
	2	433	88 (20.32)	12 (13.64)	76 (86.36)	35 (46.05)	41 (53.95)
						p=0.49	
	1+2	1,020	238 (23.33)	42 (17.65)	196 (82.35)	96 (48.98)	100 (51.02)
						p=0.78	
5	1	390	103 (26.41)	24 (23.30)	79 (76.70)	36 (45.57)	43 (54.43)
						p=0.43	
	2	438	109 (24.89)	14 (12.84)	95 (87.16)	48 (50.53)	47 (49.47)
						p=0.92	
	1+2	828	212 (25.60)	38 (17.92)	174 (82.08)	84 (48.28)	90(51.72)
						p=0.65	

radiologists: 79.0% likely (by χ^2) not to be random for Year 1, 81.0% for Year 2, and 93.0% for Years 1+2. The bias was against the same side (left) each year. This radiologist also had the smallest recall fraction (15.49%), fully one-quarter lower than the radiologist with the next lowest recall portion (20.68%). It is tempting to speculate that part of the gap in recall fraction reflected left-sided findings that went undetected by the radiologist of concern.

None of the other four radiologists had disparity likelihoods as great. The χ^2 statistic indicated that the highest single-year likelihood that any one of their observed disparities was genuine bias was just 58%. The highest 2-year value was merely 40%. Furthermore, by chance in this small group, it so happened that each of these other four radiologists demonstrated a small discrepancy favoring one side one year, and the other side the other year, unlike the radiologist of concern.

Many studies have demonstrated a higher frequency of breast cancer on the left, while others not. Attempts to explain left-sided predominance have consistently failed. If genuine, such predominance may apply only in certain racial or ethnic groups, and the extent is, in any event, slight.[10] Furthermore, the laterality bias detected in the concerning radiologist was against the left side.

Certainly, based simply on this single radiologist in the authors' small volume practice, they do not propose that many radiologists share the same bias. On the other hand, it is possible that a small percentage of radiologists may; if so, quality improvement warrants identifying who they are in audit. Moreover, this early finding may be the basis to evaluate the matter: to research a large number of radiologists in a high-volume practice or group of practices.

An exhaustive, English-language, literature search failed to discover any discussion regarding what portion of screening recalls should be bilateral, nor how much unilateral recalls may appropriately diverge from 50–50, left–right, nor if there may exist laterality bias in any particular radiologist or group of radiologists. Only one, merely tangentially related study was found.[11] The authors' intention was to discover if the "excess of left-sided breast cancers" is due to detection being more common on the left by radiologists. In contradistinction to the current work, that study was experimental in design, cancer-enriched test cases were shown to eight radiologists (three not mammogram readers at the time), image-correlation was done to assess report validity, recalls were based only on microcalcifications (excluding masses, asymmetries, and areas of architectural distortion), and the bilateral recall fraction was not reported nor could it be gleaned from the reported data.

That study "did not detect any left- or right-sided bias in perceptual detection of microcalcifications in the reader group." It is unclear whether that was in reference to the reader group as a whole or to individuals within the group.

Laterality bias would not necessarily be expected to be detected by current metrics. A laterality bias necessary to place an individual outside the benchmarks of existing quality metrics would have to be great. Even if the fraction of radiologists who exhibit laterality bias is small, not affecting group statistics significantly, detection of individuals would be helpful toward increasing their awareness and subsequent quality. Focusing on individuals would allow bias to be discovered when earlier or mild, raising consciousness. Once detected, the cause may be investigated, discerned, and remediated.

To the authors' knowledge, visual acuity and visual field detection are not routinely tested in radiologists in most (if not all) countries. It is tempting to speculate whether there might have been a visual field deficit or perhaps decreased range of cervical motion related to degenerative change. Many factors may conceivably affect laterality bias, including (but not limited to) display arrangement,

reflections, background lighting (apart from reflections), hanging protocol, reader position in relation to others and to displays, visual field deficits, unilateral chronic neck pain, and decreased range of neck motion.

In the authors' small-volume, high-positivity or high recall-rate practice, they found the portion of screening bilateral recalls to be 16.5% (16.1% excluding the radiologist with possible laterality bias). The authors speculate that less confident radiologists may display a higher percentage than others; they did not assess radiologist confidence level. Since there may be a relationship between experience and confidence, more experienced radiologists might be expected also to display a lower bilateral percentage; that was not the case in this small-volume audit. Also, a higher bilateral percentage may be appropriate in a practice in which patients more often have had no prior mammogram, or the prior mammogram had been long ago. The authors propose that a range should be established, anticipating that that range might include 16%; for example, perhaps it may prove to be 10–20%.

Limitations of this analysis include the small number of radiologists considered and the relatively small volume practice. A study with a very large number of radiologists could help to confirm the existence of laterality bias in other radiologists and to set benchmarks for the proposed metrics.

A study with very large numbers of mammograms and of radiologists would also increase statistical power. In that regard, an alpha of 0.05 is usually utilized in medical statistics as the threshold of significance: if the p-value is less than the chosen alpha, the result is arbitrarily labelled 'significant'. As professional, medical statisticians remind us, "Many current research articles specify an alpha of 0.05 for their significance level. It cannot be stated strongly enough that there is nothing special, mathematical, or certain about picking an alpha of 0.05."[12] Setting the alpha at 0.05, albeit common, is wholly arbitrary. True findings may exist at a p-value of 0.08, and false findings may exist at a p-value of 0.02. Hence, the value of research to be done based on this preliminary, concerning finding, with a larger group of radiologists reading a larger volume of mammograms, to increase statistical power.

Exclusion of BI-RADS 3, 4, and 5 categorizations may perhaps seem a limitation of this audit, but it was not. As a matter of practice on quality and operational grounds, potentially BI-RADS 3, 4, and 5 findings at screening mammography were (and still are) preliminarily categorized as BI-RADS 0; all such cases were given BI-RADS categorization 0, and were included in the BI-RADS 0 reports tallied.

Unilateral screening mammograms are often interpreted with a different hanging protocol from bilateral exams. In order not to introduce the extraneous variable of hanging protocol, unilateral mammograms were excluded from this study. It could prove informative to compare unilateral left versus right exams since their hanging protocols would presumably be the same. For instance, perhaps there is no laterality bias in the setting of viewing just one breast; that is, perhaps laterality bias only appears when comparing breasts side-by-side.

The authors' recall fractions were higher than those generally reported as desirable, typically approximately 5–12%.[6,13] The vast majority of the authors' practice reflects poor patients, underinsured or altogether uninsured, and mostly immigrants who may not trust governmental institutions like the authors', may not speak English, and are new or relatively new to healthcare in the USA. For those and related reasons, a strikingly high portion of the authors' patients have had no prior mammogram; or, if they have, it often was not in recent years, not obtainable, and/or of poor quality. When women had no prior mammogram or the most recent mammogram was more than 3 years old, one study found screening recall rates were 67% higher.[14] Since comparison to prior exams of quality, particularly those over the past several

years, often obviates recall, the high average recall amongst the authors' five radiologists is understandable. Furthermore, in a large, multicenter report, four of thirteen sites had recall rates with mammograms done, including tomosynthesis, that were well above the recommended rates for digital mammography without tomosynthesis.[15] The Breast Cancer Surveillance Consortium reports that 10% of their 359 radiologists had recall rates over 18%, with some approaching 30%,[12] whose patients may face similar impediments to those faced by the authors.

There is no substantive limit to the generalizability of this discovery. Implementation of a tally process regarding laterality and bilaterality would not be envisioned to be difficult, costly, or very time-consuming. The impact of this observation could be substantial, particularly perhaps in detecting an outlier-performing radiologist in a large group of radiologists, yet remains to be determined.

CONCLUSION

Laterality bias may exist in a radiologist who interprets screening mammograms, reflected by rate of advising left versus right immediate recall. The portion of reports recommending recall that is bilateral may simultaneously be assessed. Laterality and bilaterality biases could conceivably occur in the same reader. The authors do not, based simply on their small volume audit, propose what these values should, with high quality, be. How far unilateral recall recommendations may, with high quality, diverge from 50–50, left–right, and what a high-quality range of bilateral (versus unilateral) recalls is or should be, both have the potential to become valuable, quality metrics in screening mammography. These concerns have never before been discussed, let alone addressed. The authors call for them to be evaluated further.

References

1. Feig SA. Auditing and benchmarks in screening and diagnostic mammography. Radiol Clin North Am. 2007:45(5):791-800.
2. Sprague BL et al. National performance benchmarks for modern diagnostic digital mammography: update from the Breast Cancer Surveillance Consortium. Radiology. 2017;283(1):59-69.
3. PenRad Technologies, Inc. PenRad Mammography Information System. Available at http://www.penrad.comcastbiz.net/pdfs/mainsales_new_address_2715.pdf. Last accessed: 20 June 2024.
4. Sprague BL et al. New mammography screening performance metrics based on the entire screening episode. Cancer. 2020;126(14):3289-96.
5. Sickles EA et al., "ACR BI-RADS® Mammography," D'Orsi CJ et al (eds.), ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System (2013), Reston: American College of Radiology; 39-48.
6. Lee CS et al. Assessing the recall rate for screening mammography: comparing the Medicare Hospital Compare dataset with the National Mammography database. AJR Am J Roentgenol. 2018;211(1):127-32.
7. D'Orsi CJ. The clinically relevant breast imaging audit. J Breast Imaging. 2020;2(1):2-6.
8. FDA. Mammography Quality Standards Act Regulations. Available at: https://www.fda.gov/radiation-emitting-products/regulations-mqsa/mammography-quality-standards-act-regulations. Last accessed: 20 June 2024.
9. U.S. Department of Health and Human Services. Code of Federal Regulations, Title 45. Available at: https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/revised-common-rule-regulatory-text/index.html. Last accessed: 20 June 2024.
10. Hennessey S et al. Bilateral symmetry of breast tissue composition by magnetic resonance in young women and adults. Cancer Causes Control. 2014;25(4):491-7.
11. Tan SY et al. Comparison of readers' detection of right-sided and left-sided breast cancers and microcalcifications. J Med Imaging Radiat Oncol. 2011;55(4):353-61.
12. Tenny S, Abdelgawad I, Statistical Significance [Internet] (2023) Treasure Island: StatPearls Publishing. Available at: https://www.ncbi.nlm.nih.gov/books/NBK459346/. Last accessed: June 2024.
13. Lehman CD et al. National performance benchmarks for modern screening digital mammography: update from the Breast Cancer Surveillance Consortium. Radiology. 2017;283(1):49-58.
14. Yankaskas BC et al. Association of recall rates with sensitivity and positive predictive values of screening mammography. AJR Am J Roentgenol. 2001;177(3):543-9.
15. Friedewald SM et al. Breast cancer screening using tomosynthesis in combination with digital mammography. JAMA. 2014;311(24):2499-507.