

Association between Radiologists' Experience and Inter-radiologist Agreement in the Final Assessment of Digital Mammography

Hye-Mi Jo¹, Seunghoon Song¹, Eun Hye Lee², Kyoungan Ko³, Bong Joo Kang⁴, Joo Hee Cha⁵,
Ann Yi⁶, Hae Kyoung Jung⁷, Jae Kwan Jun¹

¹National Cancer Control Institute, National Cancer Center

²Department of Radiology, Soonchunhyang University College of Medicine, Bucheon Hospital

³Department of Radiology, Center for Breast Cancer, National Cancer Center Hospital

⁴Department of Radiology, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea

⁵Department of Radiology and Research Institute of Radiology, Asan Medical Center, University of Ulsan,
College of Medicine

⁶Department of Radiology, Seoul National University Hospital

⁷Department of Radiology, Bundang CHA Medical Hospital, CHA University

Purpose: To quantify inter-radiologist variability in interpreting screening mammograms and analyze the experience-related factors affecting inter-radiologist agreement.

Materials and Methods: A random sample of 300 mammograms was selected from women aged ≥ 40 years who were examined under the National Cancer Screening Program (NCSP) in Korea from 2007 to 2009. Inter-radiologist agreement was assessed with a weighted kappa statistic (K_w). The factors affecting the radiologists' agreement were assessed using logistic regression.

Results: The inter-radiologist agreement was moderate (overall $K = 0.60$, 95% confidence interval [CI]: 0.44–0.78) regarding the decision to recall vs. no recall and substantial (overall $K_w = 0.64$, 95% CI: 0.57–0.71) using the four-category scales for final assessment in the NCSP. Radiologists who read more than 5000 screening mammograms annually had significantly higher agreement (OR, 1.36; 95% CI, 0.95–1.95), compared to those who did not. Similarly, those who spent ≥ 6 hours weekly reading screening mammograms had significantly higher agreement (OR, 1.76; 95% CI, 1.34–2.31), compared to those who spent less time.

Conclusion: Inter-radiologist agreement varied widely. However, radiologists who read a higher annual volume of screening mammography and spent more time per week doing so demonstrated higher agreement in the final assessment.

Index words: Mammography; Breast cancer; Observer variation; Reading volume

Introduction

Breast cancer is the second most common malignancy among women in Korea, accounting for 17.6% of all cancers in Korean women in 2014 (1). The highest incidence of breast cancer is found among Korean women in their forties with relatively high breast densities and decreases after age 50, whereas breast cancer incidence in the United States continues to increase with age (2).

Since 1999, the National Cancer Screening Program (NCSP) in Korea has recommended biannual screening mammograms for women aged 40 years and older (3). Breast cancer screening by mammography is one of the most widely used methods and is the only evidence-based screening procedure currently available to reduce breast cancer mortality. However, there is wide variability in mammogram interpretation among radiologists (4–6), which can lead to substantial differences in their recommendations for follow-up. Additionally, this variability in mammography reading may affect the quality of screening programs, causing undetected tumors, diagnostic delay, false positives, and increased costs (7, 8).

This variability has been suggested to be associated with experience-related factors such as annual reading volume (9–11). However, insufficient information is currently available regarding the experience-related factors affecting inter-radiologist variability in the final assessment of mammograms. The objective of this study, therefore, is to assess the inter-radiologist variability in the final assessment of screening mammography and analyze how experience-related factors affect their agreement.

Material and Methods

The NCSP collects routine medical and health data. Therefore, informed consent for this study was not required. With permission from the Ministry

of Health and Welfare, the authors used the data and radiologic images from the NCSP database. This study was approved by the Institutional Review Board (IRB) of the corresponding author's institution (approval number, NCC2014–0065).

Data Collection

This study was based on the Korean Mammographic Density (KoMAD) study, a nationwide cross-sectional survey that investigated the prevalence of dense breasts, and the association between breast density and risk of breast cancer among Korean women aged ≥ 40 years. The details of the survey are fully described elsewhere (12). From among the participants of the KoMAD study, a total of 300 were used to create the stratified random sample in the present study. All the mammograms were acquired using the standard two-view method (i.e., mediolateral-oblique and craniocaudal), with a digital mammography system.

Our study collected data about radiologists' experience in mammography reading from self-administered questionnaires. Data included years (after board certification) of experience in reading diagnostic and screening mammography results, time spent in reading mammography results daily and weekly, and annual volume of mammography results read.

The results of screening mammography in the NCSP were defined as follows: negative result (BI-RADS category I); benign (category II); incomplete, additional evaluation needed (category 0); suspicion of malignancy (categories IV and V, respectively) (13). Additionally, to compare the NCSP assessment of recall versus no recall, positive results (incomplete, additional evaluation needed or suspicion of malignancy) were grouped as “recall”, and negative results (negative and benign) as “no recall”.

Data Analysis

We assessed the inter-radiologist agreement in the final assessment for reporting breast cancer

screening results (recall vs. no recall) using the percentage of agreement, as well as Cohen's kappa statistic (K) and its 95% confidence interval (CI). Because the results of screening mammography in the NCSP categories are an ordinal scale, we also used the weighted kappa statistic (K_w). Kappa values were interpreted as follows: poor, ≤ 0.0 ; slight, >0.0 to 0.2 ; fair, >0.2 to 0.4 ; moderate, >0.4 to 0.6 ; substantial, >0.6 to 0.8 ; and almost perfect, >0.8 to 1.0 (14). Because K can be used only for comparisons between two radiologists, we estimated an overall kappa value for multiple radiologists and multiple categories using Krippendorff's alpha, which is equivalent to an overall weighted kappa (15).

We assessed the effects of the radiologists' experience-related factors on inter-radiologist agreement in the final assessment using a logistic regression model with odds ratio (OR) and 95% CI. The dependent variable was an indicator of agreement for which 1 and 0 denoted 'agree' and 'disagree', respectively. The associations between trends in outcomes and categorical factors were tested using categorical factors entered in the model as continuous variables. Values of $P < 0.05$ were considered statistically significant, and all statistical analyses were conducted using the SAS software, version 9.2 (SAS Institute, Cary, NC).

Results

The participating radiologists' characteristics are summarized in Table 1. The radiologists had a mean of 7.5 years of experience in reading mammograms. Five radiologists had undergone fellowship training as a breast specialist at general hospitals. They reported spending an average of 5.3 hours a week reading screening mammograms, and had interpreted an average of 4,800 screening mammograms per year during the previous 5 years. The findings of the six radiologists regarding the need for recall and breast density are summarized in Table 2.

The results of the statistical analysis of inter-radiologist agreement are shown in Table 3. Regarding the need for recall, the K values (95% CI) ranged from 0.49 (0.39–0.59) to 0.76 (0.67–0.85), with an overall K of 0.60 (0.41–0.78). Regarding the four-category scales, K_w values ranged from 0.52 (0.44–0.61) to 0.72 (0.64–0.78), with an overall K_w of 0.64 (0.57–0.71). The highest agreement was found between radiologists D and E ($K_w = 0.72$), in their assessment of the four-category scales.

Inter-radiologist agreement differed significantly based on the population characteristics (Table 4). The inter-radiologist agreement was lower when breast density was heterogeneously dense or

Table 1. Characteristics of the Participating Radiologists

Characteristics	n
Breast specialist	
Yes	6
No	-
Fellowship training in breast or women's imaging	
Yes	5
No	1
Years of reading mammography	
< 10	4
≥ 10	2
Average number of mammograms read per year during the past 5 years*	
< 10,000	3
$\geq 10,000$	3
Average number of screening mammograms read per year during the past 5 years	
< 5,000	4
$\geq 5,000$	2
Daily time spent in reading screening mammography (%)	
< 25	3
≥ 25	3
Weekly time spent in reading screening mammography (hour)	
< 6	3
≥ 6	3

Note.-*Excluding ultrasonography and magnetic resonance imaging, but including both screening and diagnostic mammography.

extremely dense (OR, 0.49; 95% CI, 0.42–0.59). Radiologists with an annual screening reading volume of 5,000 or more mammograms were in higher agreement (OR, 1.36; 95% CI, 0.95–1.95), compared to those who did not. Similarly,

radiologists who spent more than 6 hours of weekly working time in reading screening mammograms showed a 76% higher agreement (OR, 1.76; 95% CI, 1.34–2.31). Years of experience in reading mammograms were not significantly associated with

Table 2. Radiologists' Assessment of Breast Density and Screening Results

Parameter	% of total						Median
	Rad. A	Rad. B	Rad. C	Rad. D	Rad. E	Rad. F	
Breast density *							
Grade 1	22.9	19.4	10.8	17.4	12.5	14.6	16.0
Grade 2	32.3	25.7	29.2	35.1	21.5	30.2	29.7
Grade 3	27.3	35.4	35.8	23.6	53.5	22.9	31.4
Grade 4	17.4	19.4	24.3	24.0	12.5	32.3	21.7
Screening results							
Negative	69.8	60.4	58.7	76.0	74.7	66.0	67.9
Benign	5.6	0.3	13.9	2.4	0.7	0.0	1.6
Incomplete, additional evaluation needed	18.4	31.6	20.1	13.5	20.8	19.1	19.6
Suspicion of malignancy	6.3	7.6	7.3	8.0	3.8	14.9	7.5

Note.- *Breast Imaging Reporting and Data System (BI-RADS) breast density categories: 1 = entirely fatty, 0–25%; 2 = scattered fibroglandular tissue, 26–50%; 3 = heterogeneously dense, 51–75%; 4 = extremely dense, 76–100% (13).

Table 3. Analysis of Inter-radiologist Agreement using Cohen's Kappa Statistic (K) and Weighted Kappa (K_w)

Pairs of radiologists	No Recall vs. Recall		NCSP final assessment	
	Agreement, %	$K^{\dagger}$ (95% CI)	Agreement, %	$K_w^{\ddagger}$ (95% CI)
A,B	80.6	0.56 (0.47–0.66)	75.3	0.61 (0.51–0.69)
A,C	85.4	0.62 (0.52–0.73)	72.2	0.61 (0.53–0.70)
A,D	87.8	0.66 (0.56–0.76)	82.6	0.70 (0.62–0.78)
A,E	88.2	0.68 (0.58–0.78)	81.9	0.69 (0.61–0.77)
A,F	82.3	0.58 (0.48–0.68)	72.2	0.56 (0.47–0.65)
B,C	78.5	0.52 (0.42–0.62)	66.3	0.53 (0.44–0.61)
B,D	77.4	0.49 (0.39–0.59)	72.9	0.53 (0.44–0.63)
B,E	81.3	0.58 (0.48–0.67)	74.7	0.56 (0.46–0.64)
B,F	80.9	0.59 (0.49–0.69)	71.9	0.58 (0.50–0.66)
C,D	86.5	0.64 (0.53–0.74)	71.9	0.60 (0.52–0.68)
C,E	87.5	0.68 (0.58–0.77)	71.2	0.60 (0.52–0.67)
C,F	83.7	0.62 (0.52–0.72)	65.2	0.55 (0.47–0.63)
D,E	91.3	0.76 (0.67–0.85)	83.7	0.72 (0.64–0.78)
D,F	83.3	0.59 (0.49–0.69)	76.4	0.59 (0.51–0.69)
E,F	81.6	0.56 (0.46–0.66)	71.5	0.52 (0.44–0.61)
Overall weighted kappa [§]				
A, B, C, D, E, F	87.4	0.60 (0.41–0.78)	80.7	0.64 (0.57–0.71)

Note.- CI, confidence interval; [†]Standard K-value: patient recall recommended (incomplete or suspicion for malignancy) vs. no recall (negative and benign); [‡]weighted kappa: measurement of four categories in the NCSP; [§]Krippendorff's alpha (15).

inter-radiologist agreement.

Discussion

In this study, we investigated how radiologists' experience and mammography workload affected their agreement regarding the final assessment of digital mammograms. We found that the inter-radiologist agreement was moderate (overall $K = 0.60$) on recall vs. no recall, and substantial (overall $K_w = 0.64$) on the four categories of final assessment in the NCSP. However, inter-radiologist agreement in the final assessment varied widely. The variability

in final assessment decreased with increasing screening reading volume and amount of time spent in reading screening mammography, and increased with increasing breast density. The variability in final assessment was not significantly associated with the radiologists' experience in terms of years of reading mammograms.

A few studies (5-7, 16) have examined inter-radiologist agreement in the final assessment of mammography and reported fair to moderate inter-radiologist agreement, which was lower than that found in the present study. This discrepancy might be due to the difference between diagnostic

Table 4. Likelihood of Inter-radiologist Concordance according to Radiologists' Characteristics

Category	Odds ratio (95% CI)	P for trend
Subjects' characteristics		
Breast density*		
Fat	1.00 (Reference)	<0.001
Dense	0.49 (0.42–0.59)	
Radiologist experience		
Years of reading mammography		
Both radiologists <10	1.00	0.890
One <10 and one ≥10	1.03 (0.87–1.22)	
Both radiologists ≥10	0.98 (0.70–1.37)	
Average number of mammograms read per year during the past 5 years [†]		
Both radiologists <10,000	1.00	0.421
One <10,000 and one ≥10,000	1.02 (0.83–1.25)	
Both radiologists ≥10,000	1.14 (0.85–1.52)	
Average number of screening mammograms read per year during the past 5 years		
Both radiologists <5,000	1.00	0.038
One <5,000 and one ≥5,000	1.15 (0.97–1.36)	
Both radiologists ≥5,000	1.36 (0.95–1.95)	
Daily time spent in reading screening mammography (%)		
Both radiologists <25	1.00	0.068
One <25 and one ≥25	1.05 (0.85–1.29)	
Both radiologists ≥25	0.80 (0.62–1.03)	
Weekly time spent in reading screening mammography (hour)		
Both radiologists <6	1.00	<0.001
One <6 and one ≥6	1.13 (0.92–1.38)	
Both radiologists ≥6	1.76 (1.34–2.31)	

Note.-* Breast Imaging Reporting and Data System (BI-RADS) breast density categories: 1 = entirely fatty, 0–25%; 2 = scattered fibroglandular tissue, 26–50%; 3 = heterogeneously dense, 51–75%; 4 = extremely dense, 76–100% (13); [†]Excluding ultrasonography and magnetic resonance imaging, but including both screening and diagnostic mammography.

and screening mammograms, types of screening mammography, subjects' characteristics, and experience of participating radiologists.

The high agreement of final assessments in this study may be related to the types of equipment used in screening mammography. Other studies were based on the reading of film-based screening mammography (6, 7, 16), but this study used digital mammography, which has several advantages over film-based imaging including better resolution (17). Therefore, the use of digital mammography could be a reason for the decrease in the variability of mammography reading. Kerlikowske et al. (7) found that there was more variability in film interpretation among women with dense breasts. We also found that dense breasts increased the disagreement among radiologists. That is, high breast density impairs the sensitivity and specificity of screening mammography (9). Previous studies (5–7) included both diagnostic and screening mammograms, whereas the breast disease statuses among their study population varied greatly. In this study, inter-radiologist agreement was higher than in these studies, because we only assessed screening mammograms.

The high agreement in this study could also stem from the fact that all participating radiologists were well-trained in breast imaging and had at least 5 years of experience in mammography reading. Berg et al. (5) reported fair agreement (overall $K = 0.21$) among experienced radiologists not specifically trained in standard criteria, and found that disagreement was the greatest between “probably benign” and “suspicious”. Kerlikowske et al. (7) reported moderate agreement (overall $K = 0.58$) among experienced radiologists interpreting screening mammograms, and found a decreased variability in mammography reading after training radiologists to identify breast lesions. To reduce inter-examiner variability and potentially improve the accuracy of mammography, two studies have suggested that it is necessary to provide adequate

training to radiologists, as well as to standardize the interpretation criteria. In a Korean study, Lee et al. (16) reported moderate agreement ($K = 0.44$, lower than that in the current study) among five radiologists who were breast imaging specialists and radiology residents. Since 2002 in Korea, the reading of breast imaging has become a standard component of radiology residents' training programs (18). That previous study (16) reflects the results of data before radiology fellowship programs were offered. According to Lee et al., the inter-radiologist variability was affected by the lack of specific training and indication standards.

However, our findings suggest that mammography reading still ultimately depends on the individual radiologists. In our study, radiologists who spent ≥ 6 hours of weekly working time in reading screening mammograms had a significantly higher agreement rate. In addition, the annual volume of screening mammography reading was also an important factor. Radiologists reading more than 5000 mammograms annually had a significantly higher agreement than those reading less than 5000 mammograms. This result is in agreement with previous findings that an increased mammography workload is associated with improved reading performance among radiologists (19–22). The Mammography Quality Standards Act in the USA (19) requires radiologists to interpret a minimum of 480 mammograms per year to improve performance, while other countries set much higher requirements, such as a minimum of 5,000 in the United Kingdom and 2,000 in the Canada (20). Two previous studies also found an association between the volume of screening mammograms interpreted and overall agreement.

There are a few limitations in this study. First, we compared the agreement in final assessment for digital mammography only, and not for screen-film and computed radiography mammography. However, digital mammography may be a more precise and consistent means of evaluation. Second, our study did not measure intra-radiologist

variability. It was not possible to perform this analysis because of limited data. Third, only specialists in breast imaging participated in the study, which means that our findings may not be generalizable to general radiologists, particularly those who only interpret a low volume of screening mammograms. Despite these limitations, this study was able to analyze how the years of experience and workload in screening mammography affected variability in the final assessment.

Conclusion

Inter-radiologist agreement was moderate to substantial on the final assessment in the NCSP. Working time in reading mammography and annual reading volume of screening mammograms affected the inter-radiologist variability in the final assessment on digital mammography. Finally, programs should be implemented to allow radiologists to spend more time practicing mammography reading to decrease inter-radiologist variability.

Acknowledgements

This study was supported by the National Cancer Center of Korea (NCC) Grant (no. 1610402-2).

References

1. Jung KW, Won YJ, Oh CM, Kong HJ, Lee DH, Lee KH. Cancer statistics in Korea: incidence, mortality, survival, and prevalence in 2014. *Cancer Res Treat* 2017;49(2):292-305
2. Park SJ, Kang D, Kim YJ, Yoo KY. Epidemiologic characteristics of the breast cancer in Korea. *J Korean Med Assoc* 2009;52:937-45
3. Suh M, Choi KS, Park B, et al. Trends in cancer screening rates among Korean men and women: results of the Korean National Cancer Screening Survey, 2004-2013. *Cancer Res Treat* 2016;48(1):1-10
4. Carney PA, Elmore JG, Abraham LA, et al. Radiologist uncertainty and the interpretation of screening. *Med Decis Making* 2004;24(3):255-264
5. Berg WA, Campassi C, Langenberg P, Sexton MJ. Breast Imaging Reporting and Data System: inter- and intraobserver variability in feature analysis and final assessment. *AJR Am J Roentgenol* 2000;174(6):1769-77
6. Redondo A, Comas M, Macia F, et al. Inter- and intraradiologist variability in the BI-RADS assessment and breast density categories for screening mammograms. *Br J Radiol* 2012;85(1019):1465-70
7. Kerlikowske K, Grady D, Barclay J, et al. Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System. *J Natl Cancer Inst* 1998;90(23):1801-9
8. Elmore JG, Miglioretti DL, Reisch LM, et al. Screening mammograms by community radiologists: variability in false-positive rates. *J Natl Cancer Inst* 2002;94(18):1373-80
9. Barlow WE, Chi C, Carney PA, et al. Accuracy of screening mammography interpretation by characteristics of radiologists. *J Natl Cancer Inst* 2004;96(24):1840-50
10. Theberge I, Hebert-Croteau N, Langlois A, Major D, Brisson J. Volume of screening mammography and performance in the Quebec population-based Breast Cancer Screening Program. *CMAJ* 2005;172(2):195-9
11. Molins E, Macia F, Ferrer F, Maristany MT, Castells X. Association between radiologists' experience and accuracy in interpreting screening mammograms. *BMC Health Serv Res* 2008;8:91
12. Jun JK, Kim MJ, Choi KS, Suh M, Jung KW. Development of a sampling strategy and sample size calculation to estimate the distribution of mammographic breast density in Korean women. *Asian Pac J Cancer Prev* 2012;13(9):4661-4
13. American College of Radiology. BI-RADS® breast imaging and reporting data system: breast imaging atlas, 4th ed. Reston, VA, 2003
14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33(1):159-74
15. Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. *Commun Methods Meas* 2007;1:77-89
16. Lee KJ, Lee WC, Hwang IY, et al. Interobserver variability in interpretation of mammogram. *J Korean*

- Radiol Soc 2004;51:351-6
17. Séradour B, Heid P, Estève J. Comparison of direct digital mammography, computed radiography, and film-screen in the French national breast cancer screening program. *AJR Am J Roentgenol* 2014;202(1):229-36
 18. Lee EH, Jun JK, Jung SE, Kim YM, Choi N. The efficacy of mammography boot camp to improve the performance of radiologists. *Korean J Radiol* 2014;15(5):578-85
 19. Rickard M, Taylor R, Page A, Estoesta J. Cancer detection and mammogram volume of radiologists in a population-based screening programme. *Breast* 2006;15(1):39-43
 20. Buist DS, Anderson ML, Haneuse SJ, et al. Influence of annual interpretive volume on screening mammography performance in the United States. *Radiology* 2011;259(1):72-84
 21. Huo CW, Chew GL, Britt KL, et al. Mammographic density—a review on the current understanding of its association with breast cancer. *Breast Cancer Res Treat* 2014;144(3):479-502
 22. Theberge I, Chang SL, Vandal N, et al. Radiologist interpretive volume and breast cancer screening accuracy in a Canadian organized screening program. *J Natl Cancer Inst* 2014;106(3):461

대한유방검진학회지 2017;14:94-101

선별유방촬영술의 판독자간 일치도 분석: 일치도에 영향을 미치는 판독자 요인

조혜미¹ · 송승훈¹ · 이은혜² · 고정란³ · 강봉주⁴ · 차주희⁵ · 이 안⁶ · 정혜경⁷ · 전재관¹

¹국립암센터 국가암관리사업본부 암정보교육과, ²순천향대학교 의과대학 부천병원 영상의학과,
³국립암센터병원 유방암센터 영상의학과, ⁴가톨릭대학교 의과대학 서울성모병원 영상의학과,
⁵울산대학교 의과대학 서울아산병원 영상의학과, ⁶서울대학교 의과대학 서울대학교병원 영상의학과,
⁷차의과대학교 분당차병원 영상의학과

목적: 유방암 조기검진 방법으로 널리 이용되고 있는 유방촬영술의 최종결론에 대한 판독자들 간의 일치율을 평가하고, 이를 통하여 판독자들 간의 일치율에 영향을 주는 요인들을 파악하고자 하였다.

대상과 방법: 2007-2009년 국가 암조기검진사업을 통해 유방암 검진을 받은 수검자 300명을 대상으로 6명의 판독자들로 하여금 맹검법으로 판독하도록 하였다. 판독자들 간의 일치율은 Kappa값을 이용하여 평가하였으며, 일치율에 영향을 미치는 판독자 요인을 도출하기 위하여 로지스틱 회귀분석을 시행하였다.

결과: 유방영상 최종결론에 대한 판독자간 일치도는 평균 Kappa 값이 0.64였으며, 소환률에 대한 일치도는 평균 Kappa 값이 0.60로 중등도의 일치도 값을 보였다. 로지스틱 분석을 해보면 연간 판독량과 판독시간이 증가할수록, 판독자들 간의 일치도가 유의하게 증가하였다.

결론: 판독자들 간의 일치도에는 다소 차이를 보였다. 그러나, 판독자들의 선별 유방촬영술의 대한 경험이 많을수록 일치도는 증가하였다. 유방촬영술의 판독의 차이를 줄여나가기 위해서는 판독자들의 판독 경험을 고려하여야 한다.

Index words: Mammography; Breast cancer; Observer variation; Reading volume

Corresponding author: Jae Kwan Jun, Ph.D, M.D.