Original Article

Interpretive Volume and Inter-Radiologist Agreement on Assessing Breast Density

Hye-Mi Jo[1], Seunghoon Song[1], Eun Hye Lee[2], Kyungran Ko[3], Bong Joo Kang[4], Joo Hee Cha[5], Ann Yi[6], Hae Kyoung Jung[7], Jae Kwan Jun[1]

[1]*National Cancer Control Institute, National Cancer Center, Goyang, Korea*
[2]*Department of Radiology, Soonchunhyang University College of Medicine, Bucheon Hospital, Bucheon, Korea*
[3]*Department of Radiology, Center for Breast Cancer, National Cancer Center Hospital, Goyang, Korea*
[4]*Department of Radiology, St. Paul Hospital, College of Medicine, The Catholic University of Korea, Seoul, Korea*
[5]*Department of Radiology and Research Institute of Radiology, Asan Medical Center, University of Ulsan, College of Medicine, Seoul, Korea*
[6]*Department of Radiology, Seoul National University Hospital, Seoul, Korea*
[7]*Division of Radiology, Bundang CHA Medical Hospital, CHA University, Seongnam, Korea*

Purpose: The aim of this study was to investigate inter-radiologist variability in the assessment of breast density and to examine whether radiologists' experience and interpretive volume affect the screening of mammograms.

Materials and Methods: This cross-sectional study involved six radiologists who assessed breast density independently according to the Breast Imaging Reporting and Data System (Category 1: <25% glandular; Category 2: 25%-50% glandular; Category 3: 51%-75% glandular; Category 4: >75% glandular) from 300 digital mammograms obtained from women who participated in the Korean Mammographic Density study. Inter-radiologist agreement (Category 1-4) was calculated using the weighted kappa statistic and the kappa statistic for a binary classification: fatty (Category 1 and 2) versus dense (Category 3 or 4). We analyzed the effects of the radiologists' experience and volume that they read.

Results: The kappa values for inter-radiologist agreement were 0.83 (95% confidence interval [CI], 0.80-0.86 [Category 1-4]) and 0.72 (95% CI, 0.66-0.75 [fatty versus dense]). Agreement was lower in those with <10 years of experience compared with those with more experience (odds ratio [OR], 0.57; 95% CI, 0.38-0.85). The inter-radiologist agreement was significantly associated with the amount of time spent reading (OR, 0.44; 95% CI, 0.32-0.60) and screening mammography (OR, 0.68; 95% CI, 0.52-0.89).

Correspondence to: Jae Kwan Jun, MD, PhD, National Cancer Control Institute,
National Cancer Center, 323 Ilsan-ro, Ilsandong-gu, Goyang 10408, Korea
Tel: 82-31-920-2184, Fax: 82-31-920-2929
e-mail: jkjun@ncc.re.kr

Conclusion: Radiologists with more experience and a higher volume of mammography reading demonstrated higher agreement in the assessment of breast density. To improve inter-radiologist agreement, radiologists' experience in mammography reading must be considered.

Index words: Breast density; Interpretation; Inter-observer variability; Mammography

Introduction

Breast density, which is defined as the proportion of the glandular tissue composition of the breast, is a significant risk factor for breast cancer development (1). Breast density is an important variable in the risk estimation of breast cancer and impairs the performance of mammography screening (2-4). At least 22 states in the United States have recently passed breast density notification laws requiring women to be notified of their breast density results and the potential effect on the sensitivity of the mammography screening (5). In estimating an individual's risk for breast cancer development through methods such as the Gail model, it is essential to ensure high accuracy and reliability in breast density assessment (6).

Breast density can be estimated quantitatively using fully-automated volumetric measurements, semi-quantitatively using computerized thresholding techniques, or qualitatively via visual assessments such as the Breast Imaging Reporting and Data System (BI-RADS) of the American College of Radiology (7). Despite the clinical advantages of fully-automated or semi-quantitative measurements of breast density, their use has been limited by a lack of resources. Thus, the BI-RADS is the most widely used method for breast density assessment.

However, several studies have reported significant inter-observer variability in the assessment of breast density using the BI-RADS (overall weighted kappa values of 0.43-0.77) (8-11). This variability may be affected by the radiologists' experience; however, there have been no reports describing the potential factors affecting inter-radiologist variability in the assessment of breast density. Accordingly, the aim of this study was to determine inter-radiologist agreement in the assessment of breast density and to analyze factors affecting agreement.

Material and Methods

Study population

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 development among Korean women aged 40 years and older who participated in the National Cancer Screening Program (NCSP) from 2007 to 2009. Details of the survey have been fully described in a previous report (12). Prior to re-reading every mammogram entry in the KoMAD study, 300 subjects were randomly selected using a random number generator, and their mammogram data were evaluated according to the ACR BI-RADS. All mammograms consisted of the two standard views (i.e., mediolateral-oblique and craniocaudal view) of both breasts and were obtained using digital mammography.

Six radiologists who are active members of the Korean Society of Breast Imaging (Seoul, Korea) and breast specialists in general hospitals participated in this study. Information regarding the radiologists' characteristics was obtained from self-administered questionnaires (13). Data included years (after board certification) of experience in reading diagnostic and screening mammograms, time spent in reading mammography daily and weekly, and annual volume of mammography read.

Six radiologists independently assessed the breast density of all cases, based on the BI-RADS (7): Category 1, almost entirely fatty tissue (<25% glandular); Category 2, scattered fibroglandular tissue (25%–50% glandular); Category 3, heterogeneously dense (51%–75% glandular); and Category 4, extremely dense (76%–100% glandular). To compare the analysis of data for fatty versus dense breasts, Category 1 and 2 were grouped as "fatty", and Category 3 or 4 as "dense" breasts.

The NCSP database collects routine medical and health data; therefore, the requirement for informed consent for this study was waived. With permission from the Ministry of Health and Welfare (Seoul, Korea), the authors used the collected data and mammograms. This study was approved by the Institutional Review Board of the corresponding author′s institution (approval number NCC2014-0065).

Statistical analysis

The agreement in the assessment of breast density (fatty vs. dense) between two radiologists was expressed using the Cohen kappa statistic (K) and its 95% confidence interval (CI). The weighted kappa statistic (K_w) was also used because the breast density variables (Category 1 vs. 2 vs. 3 vs. 4) are based on an ordinal scale. The kappa values were interpreted as follows: poor, > 0.01; slight, ≥0.01 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 a compaison between two radiologists, the overall kappa value was used to compare multiple radiologists and variables based on the work of Hayes and Kripendorff (15). Krippendorff′s alpha, which is equivalent to the overall weighted kappa, was used as a measure of overall agreement among the radiologists.

If both radiologists′ assessment of breast density were identical, ′agree′ and ′1′ were assigned; otherwise, ′disagree′ and ′0′ were assigned. The association between the agreement and the radiologists′ experience or workload was analyzed using a logistic regression model with odds ratio (OR) and 95% CI. The associations between trends in outcomes and categorical factors were tested using categorical factors entered in the model as continuous variables. All statistical analyses were performed using SAS version 9.2 (SAS Institute, Cary, NC); $P < 0.05$ was considered statistically significant.

Results

The results of the BI-RADS assessment are shown in Figure 1. Overall, the six radiologists categorized 16% (range, 11%–23%), 29% (22%–32%), 33% (24%–36%), and 22% (13%–32%) of the breasts as Category 1, 2, 3, and 4, respectively. The radiologists′ agreement in the assessment of breast density was substantial when using two-category (overall K= 0.72) and almost perfect when using four-category (overall K_w = 0.83) (Table 1).

The radiologists reported 5–10 years of experience in reading mammograms and reading at least 5,000 mammograms annually. Among them, three radiologists reported reading ≥ 10,000 mammograms annually. All radiologists reported reading ≥ 2,000 screening mammograms per year, except one who

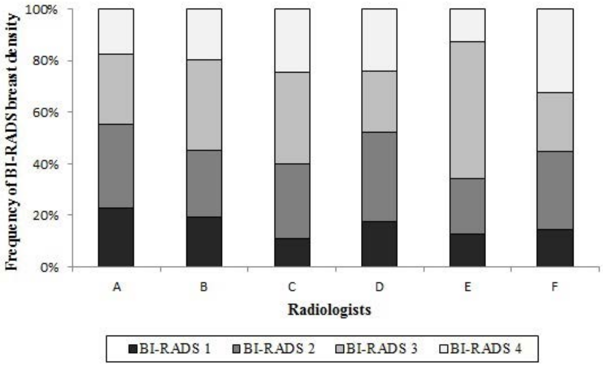

Fig. 1. Assessment of breast density using the BI-RADS (Category 1–4) by six radiologists.
BI-RADS, Breast Imaging Reporting and Data System

read diagnostic mammograms only. All but one of radiologists spent 3 h to 12 h per week reading screening mammograms (Table 2).

The years of experience and time spent in reading mammograms were significantly associated with inter-radiologist agreement (Table 3). Inter-radiologist agreement was significantly lower in those with <10 years of experience. Spending < 25% of their time reading screening mammography resulted in significantly lower inter-radiologist agreement (OR, 0.68; 95% CI, 0.52-0.89). Compared to radiologists with an annual reading volume of ≥10,000 mammograms, including both screening and diagnostic, those with a lower reading volume showed an 18% lower agreement. However, reading volume did not differ significantly in inter-radiologist agreement for assessment of breast density.

Table 1. Inter-Radiologist Agreement in the Assessment of Breast Density on Mammography

Radiologist pair	K (95% CI)†	K_W (95% CI)‡
A and B	0.75 (0.67–0.82)	0.78 (0.73–0.83)
A and C	0.65 (0.56–0.73)	0.67 (0.61–0.72)
A and D	0.82 (0.75–0.89)	0.78 (0.73–0.83)
A and E	0.55 (0.46–0.64)	0.58 (0.52–0.64)
A and F	0.74 (0.66–0.82)	0.67 (0.63–0.72)
B and C	0.74 (0.66–0.82)	0.72 (0.67–0.77)
B and D	0.80 (0.73–0.87)	0.81 (0.76–0.85)
B and E	0.70 (0.62–0.78)	0.67 (0.61–0.74)
B and F	0.81 (0.74–0.88)	0.74 (0.69–0.79)
C and D	0.71 (0.63–0.79)	0.75 (0.70–0.80)
C and E	0.70 (0.61–0.78)	0.66 (0.60–0.72)
C and F	0.72 (0.64–0.80)	0.70 (0.65–0.75)
D and E	0.62 (0.54–0.71)	0.64 (0.58–0.70)
D and F	0.81 (0.74–0.87)	0.78 (0.74–0.83)
E and F	0.68 (0.59–0.76)	0.62 (0.57–0.68)
Overall§		
A-F	0.72 (0.66–0.75)	0.83 (0.81–0.86)

†Kappa statistic: assessment of fatty (Breast Imaging and Data System [BI-RADS] Category 1 and 2) versus dense (Category 3 or 4) breast tissue.

‡Weighted kappa: measurement for Category 1 versus 2 versus 3 versus 4. §The overall kappa on using the Krippendorff's alpha (from Hayes and Krippendorff [15]).

CI, confidence interval; K, kappa statistic; K_w, weighted kappa.

Discussion

This study aimed to investigate inter-radiologist variability in the assessment of breast density and to analyze factors affecting inter-radiologist agreement. We found that the inter-radiologist agreement in the assessment of breast density based on the BI-RADS was substantial (overall K= 0.72) on two-category (fatty versus dense) and almost perfect (overall K_w = 0.83) on four-category (Category 1-4).

Table 2. Inter-Radiologist Agreement in the Assessment of Mammographic Density According to Affecting Factors

Characteristic	N	K (95% CI)†	K_W (95% CI)‡
Breast imaging experience			
Years of reading mammography			
<10	4	0.68 (0.62–0.74)	0.81 (0.78–0.84)
≥10	2	0.81 (0.74–0.87)	0.78 (0.74–0.83)
Daily time devoted to reading all mammography (%)§			
<25	2	0.55 (0.46–0.64)	0.58 (0.52–0.64)
≥25	4	0.76 (0.71–0.82)	0.86 (0.83–0.89)
Daily time devoted to reading screening mammography (%)			
<25	3	0.65 (0.58–0.72)	0.78 (0.75–0.81)
≥25	3	0.75 (0.69–0.81)	0.86 (0.84–0.89)
Weekly time devoted to reading all mammography (hours)§			
<10	1	–	–
≥10	5	0.73 (0.67–0.79)	0.83 (0.80–0.86)
Weekly time devoted to reading screening mammography (hours)			
<5	2	0.74 (0.66–0.82)	0.67 (0.63–0.72)
≥5	4	0.71 (0.65–0.77)	0.83 (0.79–0.86)
Interpretive volume			
Annual volume of mammography read§			
<10,000	3	0.66 (0.59–0.72)	0.80 (0.77–0.84)
≥10,000	3	0.74 (0.68–0.81)	0.86 (0.83–0.89)
Annual volume of screening mammography read			
<5,000	4	0.70 (0.66–0.73)	0.82 (0.79–0.84)
≥5,000	2	0.71 (0.63–0.79)	0.75 (0.70–0.80)

†Kappa statistic: measurement for fatty versus dense breast tissue.

‡Using Cohen's weighted kappa for comparison between two radiologists and Krippendorff's alpha (from Hayes and Krippendorff [15]) for multiple radiologists.

§Except for ultrasonography and magnetic resonance imaging. All mammography: both screening and diagnostic mammography.

Table 3. Odds Ratios for Inter-Radiologist Agreement in the Assessment of Mammographic Breast Density

Parameter	Odds ratio (95% CI)[†]	p for trend
Years of reading mammography		
Both radiologists, <10	0.57 (0.38–0.85)	
Radiologists, <10 and ≥ 10	0.71 (0.47–1.07)	< 0.001
Both radiologists, ≥ 10	1.00 (Reference)	
Daily time devoted to reading all mammography (%)[‡]		
Both radiologists, < 25	0.44 (0.32–0.60)	
Radiologists, < 25 and ≥ 25	0.78 (0.65–0.94)	< 0.001
Both radiologists, ≥ 25	1.00 (Reference)	
Daily time devoted to reading screening mammography (%)		
Both radiologists, < 25	0.68 (0.52–0.89)	
Radiologists, < 25 and ≥ 25	0.92 (0.73–1.16)	0.004
Both radiologists, ≥ 25	1.00 (Reference)	
Annual volume of all mammography read		
Both radiologists, <10,000	0.82 (0.61-1.10)	
Radiologists, <10,000 and ≥ 10,000	1.09 (0.85-1.42)	0.070
Both radiologists, ≥10,000	1.00 (Reference)	
Annual volume of screening mammography read[‡]		
Both radiologists, <5,000	0.98 (0.69-1.40)	
Radiologists, <5,000 and ≥ 5,000	1.11 (0.78-1.57)	0.378
Both radiologists, ≥5,000	1.00 (Reference)	

[†] The results of concordant data measurement.
[‡] Except for ultrasonography and magnetic resonance imaging. All mammography: both screening and diagnostic mammography.

This variability decreased with increasing experience (years of reading mammography) and amount of time spent in reading screening mammography. To the best of our knowledge, this study provides the first descriptive data for factors influencing inter-radiologist agreement with regard to breast density assessment.

Our study assessed breast density based on a two-category classification and a four-category classification, the results of which demonstrated substantial and almost perfect inter-radiologist agreements. However, it is possible that even when the glandular tissue volume accounts for less than 50% of the total breast volume, a high concentration of glandular tissues in only a few areas of the breast may prompt the radiologist to report the breast density as category 3 rather than category 2. Further, if glandular tissues are uniformly dispersed throughout the breast, the radiologist may report the breast as a fatty breast rather than a dense breast, even though the glandular tissue volume may account for more than 50% of the total breast volume. This suggests that density categories differ depending on each radiologist's viewpoint.

Many studies (8-11) have evaluated inter-radiologist variability in the assessment of breast density and reported moderate inter-radiologist agreement. The inter-radiologist agreement was higher in the present study compared with those reported in other countries. Berg et al. (11) reported moderate agreement in the assessment of breast density on film mammography among five radiologists (overall K_w = 0.43). Further, for radiologists not specifically trained in the BI-RADS, high breast density increased the inter-radiologist disagreement by two-fold. Ciatto et al. (8) also found moderate agreement (average K_w = 0.54) using the BI-RADS among 12 breast radiologists who read a digitized set of 100 mammograms. These radiologists were experienced in diagnostic and screening mammography (≥ 5,000 annual examinations), but had not undergone any specific training in the use of the BI-RADS density assessment. Ciatto et al. (8) described lower concordance in the inter-radiologist assessment than in the intra-radiologist assessment; this observation suggests that differences exist because breast density is perceived and interpreted differently by each radiologist. The reason for the higher agreement in the present study may be that all the participating radiologists were well trained to use the BI-RADS system. In addition, all of them had at least 5 years of experience in mammography reading. This may explain the higher inter-radiologist agreement in the present study.

Three recent studies reported findings similar to ours. Ooms et al. (9) found substantial agreement using 57 film-screen mammograms (overall K_w = 0.77) based on the four categories of breast density. Redondo et al. (10) observed substantial agreement among 21 expert radiologists using 100 film-screen mammograms (overall K_w = 0.73). Although the aforementioned studies used film-screening mammography, higher inter-radiologist agreement was achieved via oral instruction and a summary on how to apply the BI-RADS classification. Finally, Ekpo et al. (16) reported substantial agreement (overall K_w = 0.79) in full-field digital mammography among five radiologists. The reason for the good level of inter-radiologist agreement in that study was that the radiologists had recently undergone retraining in BI-RADS density assessment.

The present study examined the effects of various factors on inter-radiologist variability. We found that radiologists with more experience in mammography reading had statistically higher agreement in the assessment of breast density. In addition, radiologists who spent ≥25% of daily working time in reading screening mammograms had higher agreement. The annual volume of mammography reading was also an important factor: radiologists who read ≥10,000 mammograms annually showed higher agreement in assessment of breast density than those who did not. However, reading volume and time spent in reading mammography of radiologist were not necessarily proportional. In the case of radiologist E, the daily proportion of time spent in reading screening mammography was <10%, but the volume of screening mammograms was ≥2,500 per year. On the other hand, radiologist B spent ≥25% of their time reading screening mammograms, but volume of screening mammograms was <2,500 per year. Therefore, it can be assumed that the time spent in reading mammography was more important than the reading volume.

The current study had some limitations. First, we compared the agreement in assessment of breast density on digital mammography only (i.e., screen-film and computed radiography mammography data were excluded), because digital mammography may be a more precise and consistent means of evaluation of breast density. Second, our study did not measure intra-observer variability. It was not possible to perform this analysis because of the study design. Third, we evaluated the agreement among specialized breast radiologists only, rather than general radiologists. This may have led to overestimation of inter-radiologist agreement. Despite these limitations, the present study is one of the first to investigate how radiologists' characteristics affect agreement in the assessment of breast density using the BI-RADS.

Conclusion

Inter-radiologist agreement was substantial when rating on two-category and almost perfect on four-category. The years of experience in mammography reading and the amount of daily time spent in reading screening mammography affected the inter-radiologist agreement in the assessment of breast density on digital mammography using the BI-RADS. Future studies must consider radiologists' experience in mammography reading. In addition, improvement of the accuracy of breast density assessment may reduce variability. Thus, radiologists must continue their efforts based on specific training and clinical practice.

Acknowledgements

This research was supported by a grant of the Korea Health Technology R&D project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HC15C1257). This work was also supported by the National Cancer Center (grant number: NCC-1631130-1) and the Ministry of Health and Welfare.

References

1. Huo CW, Chew GL, Britt KL, Ingman WV, Henderson MA, Hopper JL, Thompson EW. Mammographic density-a review on the current understanding of its association with breast cancer. Breast Cancer Res Treat 2014;144:479-502
2. Drukteinis JS, Mooney BP, Flowers CI, Gatenby RA. Beyond mammography: new frontiers in breast cancer screening. Am J Med 2013;126:472-9
3. Schousboe JT, Kerlikowske K, Loh A, Cummings SR. Personalizing mammography by breast density and other risk factors for breast cancer: analysis of health benefits and cost-effectiveness. Ann Intern Med 2011;155:10-20
4. Mandelson MT, Oestreicher N, Porter PL, White D, Finder CA, Taplin SH, White E. Breast density as a predictor of mammographic detection: comparison of interval- and screen-detected cancers. J Natl Cancer Inst 2000;92:1081-7
5. Haas JS, Kaplan CP. The divide between breast density notification laws and evidence-based guidelines for breast cancer screening: legislating practice. JAMA Intern Med 2015;175:1439-40
6. Garrido-Estepa M, Ruiz-Perales F, Miranda J, et al. Evaluation of mammographic density patterns: reproducibility and concordance among scales. BMC Cancer 2010;10:485
7. ACR BI-RADS® Committee. BI-RADS® breast imaging and reporting data system: breast imaging atlas, 4th ed. American College of Radiology, Reston, 2003
8. Ciatto S, Houssami N, Apruzzese A, et al. Categorizing breast mammographic density: intra-and interobserver reproducibility of BI-RADS density. Breast 2005;14:269-75
9. Ooms EA, Zonderland HM, Eijkemans MJ, Kriege M, Mahdavian Delavary B, Burger CW, Ansink AC. Mammography: interobserver variability in breast density assessment. Breast 2007;16:568-76
10. Redondo A, Comas M, Maci à F, et al. Inter- and intraradiologist variability in the BI-RADS assessment and breast density categories for screening mammograms. Br J Radiol 2012;85:1465-70
11. 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;174:1769-77
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;13:4661-4
13. Elmore JG, Wells CK, Howard DH. Does diagnostic accuracy in mammography depend on radiologists' experience? J Womens Health 1998;7:443-9
14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74
15. Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. Communication Methods and Measures 2007;1:77-89
16. 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 Am J Roentgenol 2016;206:1119-23

디지털 유방촬영술상의 유방밀도와 판독자간 일치도 분석

조혜미[1] · 송승훈[1] · 이은혜[2] · 고경란[3] · 강봉주[4] · 차주희[5] · 이 안[6] · 정혜경[7] · 전재관[1]

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

목적: 본 연구는 유방밀도 평가에 대한 판독자간 일치도를 평가하고, 판독자의 판독경험과 판독량이 유방밀도 평가에 영향을 미치는지 조사하였다.

대상과 방법: 2007년-2009년까지 국가암조기검진사업을 통해 유방암검진을 시행한 40세 이상 여성 300명을 대상으로 하였다. 유방밀도는 전체 유방에서 선조직의 비율에 따라 분류하는 Breast Imaging Reporting and Data System (BI-RADS)를 이용하여 4가지 범주로 분류하였다. 판독자간 일치율은 Kappa값을 이용하여 BI-RADS의 4가지 범주의 일치율을 평가하였고, 일치율에 영향을 미치는 판독자 요인을 로지스틱 회귀분석을 시행하여 도출하였다.

결과: 유방밀도 평가에서 판독자간 일치도는 평균 Kappa값이 0.83였으며, 판독경험이 낮을수록 판독자들 간의 일치도가 유의하게 낮았다. 또한 판독에 소요하는 시간이 감소할수록 판독자들 간의 일치도가 유의하게 감소하였다.

결론: 판독자들의 유방촬영술의 판독경험과 판독량은 유방밀도 평가에서 판독자의 일치도를 향상시키는 중요한 요인이었다. 유방촬영술의 판독의 차이를 개선하기 위해서는 판독자의 유방촬영술의 판독경험을 고려해야 한다.

Index words: Breast density; Interpretation; Inter-observer variability; Mammography

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