

Clinical Aspects to Consider in Constructing a Standardized Test Set for Validation of Clinical Decision Support System for Mammography based on Artificial Intelligence

Jung Hyun Yoon, MD, PhD¹, Byoung Wook Choi, MD, PhD¹, Kyunghwa Han, PhD¹,
Bong Joo Kang, MD, PhD², Eun Young Ko, MD, PhD³, Hee Jung Shin, MD, PhD⁴, Jin Chung, MD, PhD⁵,
Ji-Young Hwang, MD⁶, and Eun-Kyung Kim, MD, PhD¹

¹Department of Radiology, Severance Hospital, Research Institute of Radiological Science,
Yonsei University, College of Medicine, Seoul, Korea

²Department of Radiology, Seoul St. Mary's Hospital, The Catholic University of Korea,
College of Medicine, Seoul, Korea

³Department of Radiology, Samsung Medical Center, Sungkyunkwan University,
College of Medicine, Seoul, Korea

⁴Department of Radiology, Research Institute of Radiology, Asan Medical Center, University of Ulsan,
College of Medicine, Seoul, Korea

⁵Department of Radiology, Ewha Womans University, Mokdong Hospital, College of Medicine, Seoul, Korea

⁶Department of Radiology, Kangnam Sacred Hospital, Hallym University, College of Medicine, Seoul, Korea

In the era of artificial intelligence (AI), various computed aided detection (CAD) algorithms are being developed for mammography, and the need for a standardized test set that can be generally applied for clinical validation has been voiced. During the process of constructing a standardized test set, an expert meeting was held in Seoul, Korea at June 2018, in order to reach consensus on factors to consider in building this test set for validating AI-based CAD tools developed for mammography. Faculty invited for the consensus meeting were dedicated breast radiologists, who were considered experts in this field. Topics discussed during the meeting were population-related factors, consensus on final diagnosis and interpreting lexicons according to the ACR BI-RADS, and the level of difficulty of the mammography for interpretation. Results of the discussion are summarized in this report.

In the era of artificial intelligence (AI), various computed aided detection (CAD) algorithms are being developed for mammography, and the need for a standardized test set that can be generally applied for clinical validation has been voiced. During the process of constructing a standardized test set, an expert meeting was held in Seoul, Korea at June 2018, in order to reach consensus on factors to consider in building this test set for validating AI-based CAD tools developed for mammography. Faculty invited for the consensus meeting were dedicated breast radiologists, who

Correspondence to: Eun-Kyung Kim, MD, PhD

Department of Radiology, Severance Hospital, Research Institute of Radiological Science, Yonsei University
College of Medicine, 50-1 Yonse-ro, Seodaemun-gu, Seoul 03722, Korea
Tel: 82-2-2228-7400; Fax: 82-2-393-3035, E-mail: ekkim@yuhs.ac

were considered experts in this field. Topics discussed during the meeting were population-related factors, consensus on final diagnosis and interpreting lexicons according to the ACR BI-RADS, and the level of difficulty of the mammography for interpretation. Results of the discussion are summarized in this report.

Index words: Breast; Mammography; Artificial intelligence; Validation

Introduction

Over the last century, the advance in technology has enabled access and analysis of huge amounts of accumulated data, bringing about the era of 'artificial intelligence (AI)', the branch of computer science dedicated to develop algorithms to accomplish tasks traditionally associated with human intelligence (1, 2). AI is applied on various everyday boundaries including the medical field, already substantially involved in clinical practice in areas of lesion detection, diagnosis, risk assessment or prognostication, and treatment response (3, 4). Computer-aided detection (CAD) systems had been developed and clinically used since 1966 mostly in thoracic and breast imaging (4), but has rapidly spread to multiple organs and disease entities along with the rapid development of computational resources enabling storage and processing of tremendous amount of data.

As for imaging of the breast, mammography is the primary modality in lesion visualization or detection of early changes in the breast (5), and many recent studies have reported development of CAD for mammography showing similar to higher diagnostic performances to radiologists (6–8). With the exponential development of computer technology, various machine learning/AI techniques has been applied in developing CAD algorithms for breast imaging during the recent years (9), but among them, very few been commercially utilized for clinical use, mostly due to the lack of adequate clinical validation. Prior to clinical application, external validation by applying the CAD algorithm to a study population other than one used for

training during algorithm construction is required (3). Most of the studies providing external validation are results mostly applied to a small proportion of the initial study population collected for building a CAD algorithm that has been set aside to provide validation results. Authors of this study proposed the need for a standardized test set for mammography that includes a patient cohort representing the real-world clinical environment, which would provide more integrity and generalizability to the validation results, and also enabling measurements of the strength and weaknesses of various CAD algorithms according to what it focuses on. Based on this, an expert meeting consisting of dedicated breast radiologists was organized to discuss the the factors to consider when constructing a standardized test set for validating AI-based CAD algorithms developed for mammography.

Methods

Formation of consensus group

Five radiologists who are dedicated to breast imaging with 9 to 20 years of experience (B.J.K., E.Y.K., H.J.S., C.J., and J.-Y.H.), who are also members of the Korean Society of Breast Imaging & Korean Society of Breast Screening (KSBI & KSBS) were invited to participate as faculty members of the consensus meeting. Faculty members were considered to be experts and key opinion leaders in the field, and are listed as authors.

Topics to be discussed in the consensus meeting was listed by 2 dedicated breast radiologists, each with 10 and 20 years of experience (J.H.Y., and E.-K.K.). Topics for discussion were organized into

three themes: 1) population factors to consider for inclusion in the standardized test set, 2) consensus on final diagnosis and interpreting lexicons according to the American College of Radiology Breast Imaging Reporting and Data System (ACR BI-RADS) (10), and 3) level of difficulty of the mammography for interpretation.

Consensus meeting

The consensus meeting for constructing a standardized test set for mammography CAD tools was held in Seoul, Korea on June 2018. During the meeting, an biostatistician (K.W.H) was present to provide assistance regarding the statistical factors that should be considered in validation of CAD algorithms and data required by the U.S. Department of Health Services Food and Drug Administration (FDA) (11–14).

Recommendations

1. Populations factors

1) *Collection of mammography images*

For inclusion of mammography images, the size of the centers from which the representative study population should be considered. For instance, the proportion of breast cancers among population collected from a tertiary referral center may be larger than a local clinic that people visit for their annual screening mammography. As the FDA recommends collection of consecutive cases from each participating collection site (11, 12), all experts agreed on consecutive data collection for study design. Experts agreed on collecting digital mammography images, and tomosynthesis or its synthetic images should not be included. Also, they recommended on analysis of both per patient and per breast basis.

2) *Demographics of the study population*

All experts agreed on that the standardized

test set should include mammography images of women who are ≥ 40 years, adhering to the Korean guideline for breast cancer screening which recommends screening mammography for asymptomatic women starting at the age of 40 (15). Information regarding the presence of symptoms should be collected, the proportion of symptomatic/asymptomatic groups being compensated with consecutive data collection. In addition, the experts recommended exclusion of women who had prior breast surgery, since the resultant alteration of the breast on mammography may be a confounding factor for the CAD algorithm and should be considered in a different test set.

Specific diagnosis of ductal carcinoma *in situ* (DCIS) and invasive cancers should be recorded, and if the pathologic report provides measurements for both invasive and *in situ* components, this should also be separately recorded. Size of the cancers should be correlated to the abnormal findings detected on imaging, maximum diameter of the suspicious findings seen on imaging considered as the estimated size on imaging.

3) *Mammography density*

For mammography density, the proportion of women with dense breasts has been reported to be frequently higher in Asian women than the Western population (16, 17): approximately 54.4% of Korean women aged 40 to 69 years had heterogeneously dense (grade C) or extremely dense (grade D) breasts in a recent cross-sectional study (18). The experts came into consensus that if the participating collection sites are from Korea, further data collection from the Western countries may be necessary if validation using this standardized test set should be applied to women of other ethnicity.

2. Consensus on final diagnosis and interpreting lexicons according to the ACR BI-RADS

1) Definition of breast cancer population

Definition of final diagnosis and according to its BI-RADS final assessment category is summarized in Table 1. All experts agreed that breast cancer diagnosis should include DCIS and invasive carcinomas of the breast. Cancer diagnosis should be based on pathologic diagnosis via surgery. For patients who had axilla metastasis from occult breast cancer that was not detectable on any of the breast imaging studies, mammography, ultrasonography (US), or magnetic resonance imaging (MRI), the experts agreed on excluding them from the test set.

2) Definition of benign or normal population

Experts came to consensus that normal population should consist of asymptomatic women with no abnormal findings detected on mammography and/or US or MRI obtained within 3 months of the mammography examination, assessed as

category 1 for ACR BI-RADS final assessment (10). Fat-containing lesions such as hamartomas, or low density masses such as oil cysts should be considered benign (BI-RADS category 2) without pathologic diagnosis, based on its typical mammography features (10). Typically benign calcifications assessed as BI-RADS category 2 should meet one of the features summarized in Table 1.

As for BI-RADS category 3 lesions, experts agreed in that they should be included in the study population under the following conditions; 1) confirmed as benign via biopsy, or 2) when follow up images were available and the lesion was stable for more than 2 years when not biopsied. All experts agreed in that patients who had been diagnosed to have high-risk lesions such as atypical ductal or lobular hyperplasia, lobular carcinoma in situ (LCIS) should be excluded from the population, since its risk for cancer progression or development has not yet been clarified.

Table 1. Definition of Final Diagnosis and its According ACR BI-RADS Final Assessment

	ACR BI-RADS final assessment		Definition
Negative Benign	Category 1	No detectable abnormalities on mammography and/or US, MRI	
	Category 2	Typically benign features	Masses Low-density Fat-containing Calcifications Skin Vascular Coarse or 'popcorn' like Large rod-like Round Rim Dystrophic Milk-of-calcium Suture
	Category 3	Probably benign masses showing no change or involution during $\geq$ 2years follow up	
Cancer	Regardless of mammography category	Diagnosis of DCIS or invasive carcinoma on pathology via biopsy - Exclusion of imaging-occult breast cancer with only axilla node or distant metastasis	

'ACR BI-RADS: American College of Radiology Breast Imaging Reporting And Data System, DCIS: ductal carcinoma *in situ*

3) Mammography descriptors used for analysis

All experts agreed in that the mammography descriptors used during data collection should be based on the ACR BI-RADS lexicon (Table 2) (10).

3. Level of difficulty of the mammography for interpretation

Experts came into consensus that the level of difficulty of the mammography for interpretation should be based on lesion detectability on mammography. They recommended grading the level of difficulty of each mammography case included during data collection on a 3-point scale

as follows: 1 (easily detected), 2 (visible, but not easily detected), 3 (unable to detect), scores given by dedicated breast radiologists.

Based on the consensus meeting, the case report forms (CRF) for mammography that the experts used for standardized test set is provided in Table 2.

Discussion

Only until recently has the FDA focused on the approval of AI-based healthcare devices (11–14), but the current FDA does not specifically indicate the benefits or recommended patient care by using them (19, 20). Unlike drugs that may have direct

Table 2. Case Report form (CRF) for Mammography Data

Factor	Findings
Case No.	
Age	years
Presence of symptoms	absent, palpable mass, pain, bloody nipple discharge, axilla mass
Mammography vendor	
Mammography density	Almost entirely fat (A) Scattered fibroglandular densities (B) Heterogeneously dense (C) Extremely dense (D)
Final diagnosis	Negative, benign, cancer*
Cancer detection modality	Mammography, US, MRI
Size	mm
Lesion location	Left/Right
Findings on mammography	
Mass	Shape/margin/density
Calcifications	Typically benign/suspicious morphology/distribution
Architectural distortion	
Asymmetries	Asymmetry/focal asymmetry/global asymmetry/developing asymmetry
Others	Skin lesion/intramammary lymph node/solitary dilated duct
Associated features	Skin retraction/nipple retraction/skin thickening/trabecular thickening/axillary adenopathy
Final assessment (mammography) [†]	Category 1-5
Difficulty to interpret	Easily detected (1) Visible, but not easy to detect (2) Unable to detect (3)

BI-RADS: Breast Imaging Reporting And Data System, US: ultrasonography, MRI: magnetic resonance imaging

*: DCIS and invasive cancer specified

[†]: category 0: not given, final category used in diagnostic mammography should be used for data, category 6: not given, refined into category 1-5 according to mammography features

effect on patients, approval for the use of CAD does not require high-level evidence of clinical efficacy (3, 19). The development of AI and implementation in medical imaging significantly affects patient management, and it is highly predictable that AI applications will bring about significant ethical and legal issues in healthcare in near future. Neglecting proper clinical validation of AI algorithms is a subject needing major concern, and the importance of appropriately designed external validation studies proving the efficacy of various CAD algorithms that are currently being released has been voiced (3, 19, 21). With the results of vigorous clinical validation studies, we can ensure patient benefits and well-being without having them being exposed to potential harm by using these tools.

Clinical validation, which is also termed as 'external validation', is recommended to use a separate dataset, data that were not used for algorithm building or testing which is collected from well-designed clinical cohorts to represent diagnostic performances of the AI algorithm when used in the real-world clinical practice setting (3). Authors of this study gathered in a consensus meeting to discuss factors that should be considered in a clinical cohort that may be used for external validation for AI-based CAD algorithms for mammography.

One of the major pitfalls of AI algorithms that has been acknowledged to overestimate its performance is spectrum bias, in which the patient manifestation of a specific disease in a training cohort does not reflect that of the validation cohort (3). In an effort to overcome this bias and construct a more generalized cohort representative of clinical practice, experts agreed that 1) various participation centers from local screening clinics to tertiary referral centers should be included, and 2) that data should be collected consecutively to include patients of various conditions, especially regarding demographic data such as cancer size, the presence of symptoms, the proportion of DCIS and invasive breast cancers,

mammography density, and the level of difficulty for interpretation.

As the definition for the final diagnosis has significant impact on clinical validation and also the inclusion criteria for the standardized test set, experts discussed this in depth. Breast cancer is a heterogeneous disease, in which a wide spectrum of atypia and *in situ* cancers to invasive components coexisting within a cancer mass. Therefore, experts agreed on that cancer diagnosis should be based on surgical pathology, data specifying the extent of invasive components and the *in situ* components, if available on the pathology reports. In this aspect, experts also recommended imaging-pathology correlation in regards to size measurement of breast cancer. The definition for normal or benign population was mostly based on the ACR BI-RADS (10); no abnormal findings detected on breast imaging, or benign findings that are either typical on mammography or stable during more than 2 years of follow up.

This consensus meeting has its limitations in that is that only a selection of experts were available, in which most of them are affiliated in tertiary referral centers. Clinical environment of local clinics may differ somewhat which may have not been considered during this meeting. Second, all experts were Korean, and as our standardized test set is currently considered to be applied for CAD algorithms developed in Korea, considerations regarding characteristics of other ethnicity or countries may be discussed in future international meetings.

Conclusion

This report is based on an expert meeting. Various range of topics were considered, including the study population to include in the standardized test set, consensus on final diagnosis and interpreting lexicons according to the ACR BI-RADS, and the level of difficulty for interpreting mammography.

Results of this consensus meeting provide a basis on the factors to consider when building a standardized test set that can be used for clinical validation of AI-based CAD algorithms for mammography, based on expert opinion.

Acknowledgement

This report was supported by a grant (18713MFDS331) from the Ministry of Food and Drug Safety in 2018.

References

1. Hamet P, Tremblay J. Artificial intelligence in medicine. *Metabolism* 2017;69s:S36-s40
2. Tang A, Tam R, Cadrin-Chenevert A, Guest W, Chong J, Barfett J, et al. Canadian Association of Radiologists White Paper on Artificial Intelligence in Radiology. *Can Assoc Radiol J* 2018;69:120-135
3. Park SH, Kressel HY. Connecting Technological Innovation in Artificial Intelligence to Real-world Medical Practice through Rigorous Clinical Validation: What Peer-reviewed Medical Journals Could Do. *J Korean Med Sci* 2018;33:e152
4. Giger ML. Machine Learning in Medical Imaging. *J Am Coll Radiol* 2018;15:512-520
5. Hamidinekoo A, Denton E, Rampun A, Honnor K, Zwigelaar R. Deep learning in mammography and breast histology, an overview and future trends. *Med Image Anal* 2018;47:45-67
6. Hupse R, Samulski M, Lobbes MB, Mann RM, Mus R, den Heeten GJ, et al. Computer-aided detection of masses at mammography: interactive decision support versus prompts. *Radiology* 2013;266:123-129
7. Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A. Deep Learning in Mammography: Diagnostic Accuracy of a Multipurpose Image Analysis Software in the Detection of Breast Cancer. *Invest Radiol* 2017;52:434-440
8. Samulski M, Hupse R, Boetes C, Mus RD, den Heeten GJ, Karssemeijer N. Using computer-aided detection in mammography as a decision support. *Eur Radiol* 2010;20:2323-2330
9. Yassin NIR, Omran S, El Houby EMF, Allam H. Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. *Comput Methods Programs Biomed* 2018;156:25-45
10. American College of Radiology. American College of Radiology. Breast imaging reporting and data system. 5th ed. Reston, VA: American College of Radiology, 2013
11. Guidance for Industry and FDA Staff Clinical Performance Assessment: Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Approval (PMA) and Premarket Notification [510(K)] Submissions. 2009
12. Guidance for Industry and Food and Drug Administration Staff Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Notification [510(k)] Submissions 2009
13. Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices: Guidance for Industry and Food and Drug Administration Staff 2016
14. Software as a Medical Device (SAMD): Clinical Evaluation. 2016
15. Lee EH, Park BY, Kim NS, Seo HJ, Ko KR, Min JW, et al. The Korean Guideline For Breast Cancer Screening. *J Korean Med Assoc* 2015;58
16. Rajaram N, Mariapun S, Eriksson M, Tapia J, Kwan PY, Ho WK, et al. Differences in mammographic density between Asian and Caucasian populations: a comparative analysis. *Breast Cancer Res Treat* 2017;161:353-362
17. Bae JM, Kim EH. Breast Density and Risk of Breast Cancer in Asian Women: A Meta-analysis of Observational Studies. *J Prev Med Public Health* 2016;49:367-375
18. Jo HM, Lee EH, Ko K, Kang BJ, Cha JH, Yi A, et al. Prevalence of Women with Dense Breasts in Korea: Results from a Nationwide Cross-sectional Study. *Cancer Res Treat* 2019
19. Park SH. Regulatory Approval versus Clinical Validation of Artificial Intelligence Diagnostic Tools. *Radiology* 2018;288:910-911
20. Park SH, Han K. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction. *Radiology* 2018;286:800-809

인공지능을 기반으로 한 유방촬영술용 컴퓨터 보조 진단 프로그램의 임상 검증을 위한 표준 데이터셋 구축에 고려해야 할 임상적 요인에 관한 합의 보고

윤정현¹ · 최병욱¹ · 한경화¹ · 강봉주² · 고은영³ · 신희정⁴ · 정진⁵ · 황지영⁶ · 김은경¹

¹연세대학교 의과대학 세브란스병원 영상의학과, ²가톨릭대학교 의과대학 서울성모병원 영상의학과,
³성균관대학교 의과대학 삼성서울병원 영상의학과, ⁴울산대학교 의과대학 서울아산병원 영상의학과,
⁵이화여자대학교 의과대학 이대목동병원 영상의학과, ⁶한림대학교 의과대학 강남성심병원 영상의학과

인공 지능의 시대가 도래하면서, 유방촬영술에 적용 가능한 다양한 인공지능 의료 영상 진단 보조 소프트웨어가 개발, 적용되고 있다. 이러한 인공지능 의료 진단 보조 소프트웨어를 실제 의료 환경에 적용 가능하기 위해서 이에 대한 성능 검증이 필수이고 이를 위한 표준 데이터 셋의 구축이 필요한 실정이다. 유방촬영술을 위한 표준 데이터 셋을 만드는 과정에서 어떠한 임상, 영상학적인 요소를 고려하여 만들어야 하는지에 대해 유방 영상 자문위원단을 구성하여 2018년 6월에 토론을 진행하였다. 본 토론에 참여한 자문위원단은 유방 질환 특성에 맞는 유방 영상의학 전문가들로 구성되었다. 본 토론에서 논의된 주제들은 표준 데이터 셋에 포함되는 구성원의 인구학적 요소, 최종 진단 및 판독에 사용하는 ACR BI-RADS 기술 용어에의 합의, 그리고 판독하는 유방 영상에서의 난이도 평가로 이루어졌다. 토론에서 합의된 내용들을 요약하여 보고하고자 한다.

Index words: Breast; Mammography; Artificial intelligence; Validation

Corresponding author: Eun-Kyung Kim, M.D., Ph.D.