Thieme

Artificial Intelligence for Indication of Invasive Assessment of Calcifications in Mammography Screening

Künstliche Intelligenz zur Indikationsstellung einer invasiven Mikrokalkabklärung im Mammografie-Screening

Authors
Stefanie Weigel[1], Anne-Kathrin Brehl[2], Walter Heindel[1], Laura Kerschke[3]

Affiliations
1 Clinic for Radiology and Reference Center for Mammography, University Hospital and University of Münster, Münster, Germany
2 ScreenPoint Medical, Nijmegen, The Netherlands
3 Institute of Biostatistics and Clinical Research, University of Münster, Münster, Germany

Key words
breast cancer, mammography screening, artificial intelligence, breast calcifications, positive predictive value, ductal carcinoma in situ

received 08.07.2022
accepted 16.10.2022
published online 2022

Bibliography
Fortschr Röntgenstr 2023; 195: 38–46
DOI 10.1055/a-1967-1443
ISSN 1438-9029

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Correspondence
Prof. Dr. med. Stefanie Weigel
Clinic for Radiology and Reference Center for Mammography, University Hospital and University of Münster, Albert-Schweitzer-Campus 1, Building A1, 48149 Münster, Germany
Tel.: +49/2 51/8 34 56 50
Fax: +49/2 51/8 34 56 60
weigels@uni-muenster.de

ABSTRACT

Purpose Lesion-related evaluation of the diagnostic performance of an individual artificial intelligence (AI) system to assess mamographically detected and histologically proven calcifications.

Materials and Methods This retrospective study included 634 women of one screening unit (July 2012 – June 2018) who completed the invasive assessment of calcifications. For each leasion, the AI-system calculated a score between 0 and 98. Lesions scored > 0 were classified as AI-positive. The performance of the system was evaluated based on its positive predictive value of invasive assessment (PPV3), the false-negative rate and the true-negative rate.

Results The PPV3 increased across the categories (readers: 4a: 21.2 %, 4b: 57.7 %, 5: 100 %, overall 30.3 %; AI: 4a: 20.8 %, 4b: 57.8 %, 5: 100 %, overall: 30.7 %). The AI system yielded a false-negative rate of 7.2 % (95 %-CI: 4.3 %: 11.4 %) and a true-negative rate of 9.1 % (95 %-CI: 6.6 %; 11.9 %). These rates were highest in category 4a, 12.5 % and 10.4 % retrospectively. The lowest median AI score was observed for benign lesions (61, interquartile range (IQR): 45–74). Invasive cancers yielded the highest median AI score (81, IQR: 64–86). Median AI scores for ductal carcinoma in situ were: 74 (IQR: 63–84) for low grade, 70 (IQR: 52–79) for intermediate grade and 74 (IQR: 66–83) for high grade.

Conclusion At the lowest threshold, the AI system yielded calcification-related PPV3 values that increased across categories, similar as seen in human evaluation. The strongest loss in AI-based breast cancer detection was observed for invasively assessed calcifications with the lowest suspicion of malignancy, yet with a comparable decrease in the false-positive rate. An AI-score based stratification of malignant lesions could not be determined.

Key Points:
- The AI-based PPV3 for calcifications is comparable to human assessment.
- AI showed a lower detection performance of screen-positive and screen-negative lesions in category 4a.
- Histological subgroups could not be discriminated by AI scores.

Citation Format
- Weigel S, Brehl AK, Heindel W et al. Artificial Intelligence for Indication of Invasive Assessment of Calcifications in Mammography Screening. Fortschr Röntgenstr 2023; 195: 38–46

ZUSAMMENFASSUNG

Ziel Läsionsbezogene Überprüfung der diagnostischen Wertigkeit eines individuellen Algorithmus künstlicher Intelligenz (KI) in der Dignitätsbewertung von mammografisch detektierten und histologisch abgeklärten Mikroverkalkungen.

Material und Methoden Die retrospektive Studie umfasste 634 Frauen mit abgeschlossener invasiver Abklärungsdiagnostik aufgrund von Mikroverkalkungen einer Mammografie-Screening-Einheit (Juli 2012 – Juni 2018). Das KI-System berechnete für jede Läsion einen Score zwischen 0 und 98. Scores > 0 wurden als KI-positiv betrachtet. Die KI-Performance wurde läsionen-spezifisch auf Basis des positiven prädiktiven Werts der umgesetzten invasiven Abklärungsdiagnostik (PPV3), der Rate falsch negativer und richtig negativer KI-Bewertungen evaluiert.

Ergebnisse Der PPV3 stieg über die Befundstufen an (Befunder: 4a: 21,2 %, 4b: 57,7 %, 5: 100 %, gesamt 30,3 %; KI: 4a: 20,8 %, 4b: 57,8 %, 5: 100 %, gesamt: 30,7 %). Die Rate falsch negativer KI-Bewertungen lag bei 7,2 % (95 %-CI: 4,3 %, 11,4 %), die Rate richtig negativer KI-Bewertungen bei 9,1 % (95 %-CI: 6,6 %, 11,9 %). Diese Raten waren mit 12,5 % bzw. 10,4 % in der Befundstufe 4a am größten. Im Median war der KI-Score für benigne Läsionen am geringsten (61, Interquartilsabstand [IQR]: 45–74) und für invasive Mammakarzinome am höchsten (81, IQR: 64–86). Mediane Scores für das duktale Carcinoma in situ waren: 74 beim geringen (IQR: 63–84), 70 (IQR: 52–79) beim intermediären und 74 (IQR: 66–83) beim hohen Kernmalignitätsgrad.

Schlussfolgerung Bei niedrigster Schwelle führt die Mikrokalk-bezogene KI-Bewertung zu einem zur menschlichen Bewertung vergleichbaren Anstieg des PPV3 über die Befundstufen. Der größte KI-bezogene Verlust an Brustkrebsdetektionen liegt bei geringstgradig suspekten Mikroverkalkungen vor mit einer vergleichbaren Einsparung falsch positiver invasiver Abklärungen. Eine Score-bezogene Stratifizierung maligner Läsionen lässt sich nicht ableiten.

Introduction

Mammography screening is considered the only evidence-based method for early detection of breast cancer and has been established nationwide in Germany with scientifically proven effects [1–3].

Artificial intelligence (AI) uses different algorithms to solve various tasks and can simplify tasks or support human activity [4]. Evolution of computer-aided detection (CAD) systems resulting from technical advances and deep learning algorithms can increase the efficacy of mammographic screening. A meta-analysis of retrospective mammographic studies shows that the sole application of AI algorithms can reduce the radiologists' reading workload by 17 %–91 % with a reduction in breast cancer detection of 0 %–7 % [5].

In the 50–69 years age group, calcifications constitute the second most frequent mammographic abnormality leading to further assessment and is also the second most frequent mammographic abnormality in the detection of breast cancer [6, 7]. Calcifications constitute a broad spectrum of lesions, ranging from mastopathic breast lesions to high-risk lesions and precursor lesions of invasive breast cancer to invasive cancer with varying biological significance and differing positive predictive value in invasive work-up procedures (PPV3) [8, 9].

From a user perspective, evaluating the diagnostic value of an individual AI algorithm towards calcifications is essential in order to safely integrate the abstract AI information on a defined mammographic lesion into the final human decision-making process. Retrospective integration of an available AI system [10] into the consensus conference decision-making process had the potential to mitigate false-positive recalls, but sensitivity was lower for calcifications than for masses [11].

The aim of the present retrospective study was to assess the diagnostic value of an individual AI algorithm in assessing the probability of malignancy of screen-detected, histologically-clarified calcifications with respect to human findings.

Materials and Methods

The study included 634 women with calcification-related invasive work-up after participation in a mammography-based screening program between July 2012 and June 2018. All mammographic examinations were carried oud in one screening unit. Histologically-clarified calcifications were retrospectively evaluated by an AI system at the lesion level. The evaluation of the system was compared to human evaluation and categorized based on histology, which was used as gold standard.

The study was carried out in the EU Project INTERREG V A, InMediValue 122207. Approval was obtained from the Ethics Committee of the Medical Association of Westphalia-Lippe and the Medical Faculty of the University of Münster, which had no reservations of ethical or legal nature regarding the performance of the research project.

Screening Process

As part of the German Mammography Screening Program, women between the ages of 50 and 69 are invited by letter for a two-view digital mammography screening examination. Screening mammograms are evaluated by two independent certified readers. In case of at least one abnormality, both readers discuss the case in a consensus conference together with the responsible physician, who finally decides whether a recall for further assessment is indicated and eventually performs subsequent diagnostics including invasive assessment [1].

Clinical Study Data

Grading of suspicion of malignancy in the consensus conference (4a, 4b, 5) was conducted in the screening software MaSc (KVWL, Dortmund, Germany). Categories were based on the Breast Imaging Reporting and Data System (BI-RADS) version 4 [8].

The two-view mammography was obtained at two locations (Sectra MDM L30, Linköping, Sweden; Philips MDM L50, Philips Healthcare, Eindhoven, Netherlands; Hologic 3Dimensions, Marlborough, MA, USA; Mammomat Inspiration, Mammomat Revelation, Siemens Healthcare, Erlangen, Germany). Independent double reading was performed by five readers, including two physicians supervising the program. Standardized assessment of calcification-associated lesions included sonography to exclude associated masses (Acuson S2000, Siemens Healthcare, Erlangen, Germany) and magnification views in cranio-caudal and lateral projection (Hologic Selenia Dimensions, Marlborough, MA, USA). For suspicious calcifications without any other associated findings, radiographic vacuum-assisted biopsy (Hologic Multicare Platinum, Marlborough, MA, US) was indicated as first-line method.

Data Collection

The CE- and FDA-certified AI-based software Transpara (version 1.7.0) from ScreenPoint Medical, Netherlands is a deep learning algorithm based on a deep convolutional neural network. The algorithm was trained using image data from over two million histologically-confirmed lesions and underwent external clinical validations [10]. The AI-software assigned a lesion score betweeen 1 and 100 to each suspicious region and indicated whether a calcification or mass was detected. All calcifications had been confirmed by the responsible physician based on the screening mammogram and clinical documentation. The morphology and distribution of calcifications were determined [8]. If not displayed automatically, the lesion-specific AI score was collected by user selection with a mouse click. A value of 100 represented the highest level of malignancy [12]. No score was displayed to the user (analyzed as score = 0) for lesions evaluated by the system with a score ≤ 28. For lesions rated 98–100, a score of 98 was displayed. In the case of varying scores of a lesion in the different mammographic views, the higher score was used.

Inclusion/Exclusion Criteria

Only screen-detected, calcification-associated lesions with biopsy-confirmed histological pathology were selected for the study. In case further invasive diagnostics had been recommended, all diagnostic procedures had to be completed before study inclusion. Cases with negative outcome and no indication of breast cancer were referred to biennial mammography follow-up. Exclusion criteria are shown in ▶ **Fig. 1**.

Screening-positive Calcifications

Screening-positive lesions included ductal carcinoma in situ (DCIS) and invasive breast cancer. The final surgical result was used for classification. For neoadjuvant therapy, the result of minimally invasive assessment was used. Breast cancer cases were differentiated by grade.

Screening-negative Calcifications

Histologically-benign lesions were considered screening-negative. In line with the screening evaluation, lesions of uncertain malignant potential (high-risk lesions) were also scored screening-negative. The post-surgical final histological result was used in case additional diagnostic excision was indicated. Indications for diagnostic excision were present in any case of atypical epithelial proliferation of the ductal type, as well as in residual lesion portions of flat epithelial atypia (FEA), papilloma and radial scars [13]. Screening-negative assessments were confirmed only if a two-year negative follow-up examination was present.

AI-negative and positive Calcifications

Lesions for which the AI system did not display a score (score = 0) were considered AI-negative. All lesions with a region-specific score (score ≥ 29) were considered AI-positive.

Statistical Evaluation

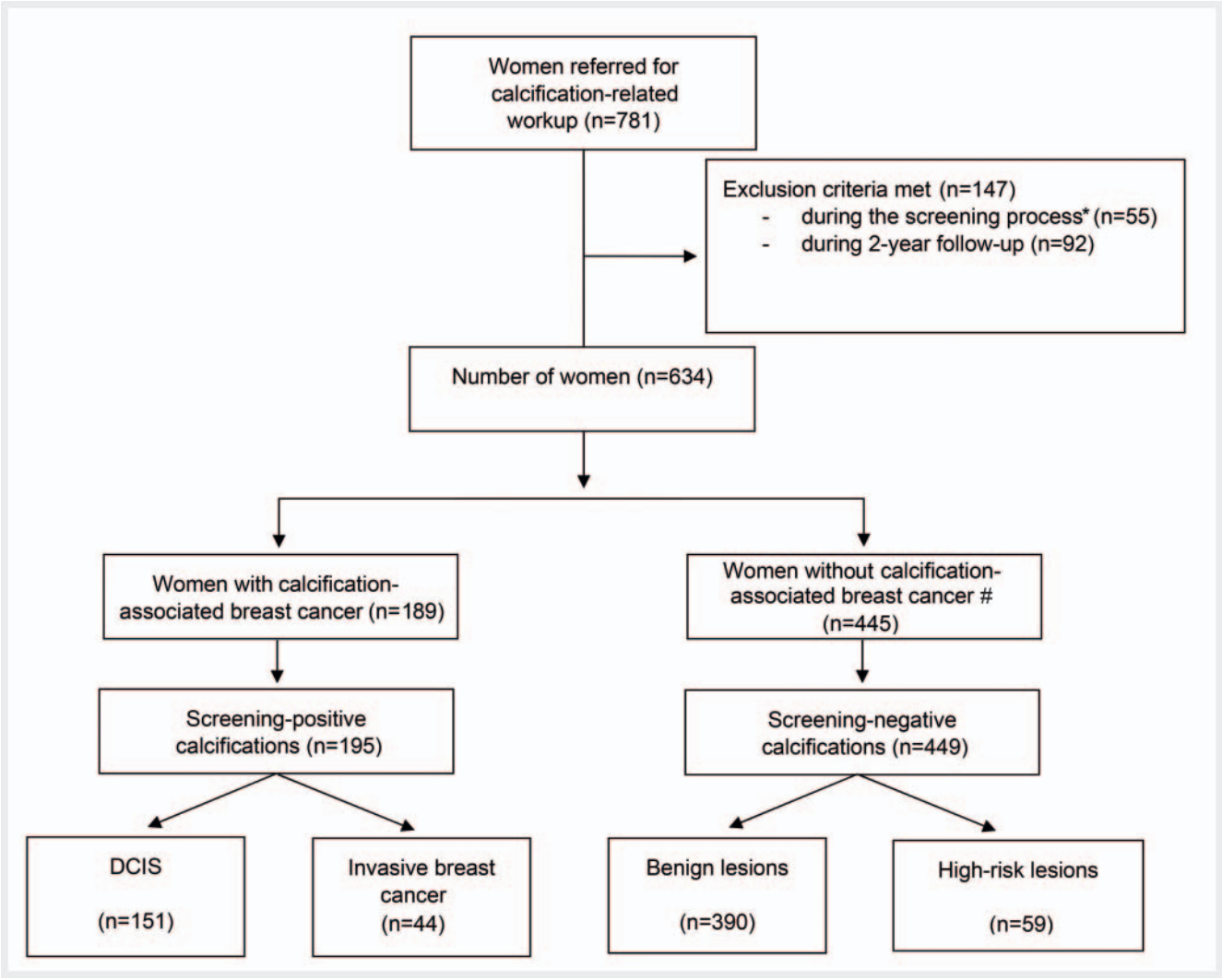
Analyses were performed using the statistics software R (version 4.0.2). Categorical parameters were presented as absolute and relative frequencies, continuous parameters were displayed as median and interquartile range. We determined the lesion-specific positive predictive value of performed invasive assessment (PPV3) towards calcifications. The performance of the AI system was evaluated based on the lesion-specific rate of false-negative evaluations, i. e., the proportion of AI-negative lesions among screening-positive calcifications (1 – sensitivity), and the rate of AI-true-negative assessments, i. e. the proportion of AI-negative lesions among screening-negative calcifications (specificity). A 95 % confidence interval was calculated for each performance indicator using non-parametric bootstrapping.

Results

Screening results

Histological results from 634 women with 644 calcifications-related histologic lesions were included (▶ **Fig. 1**). Two women received invasive assessment of calcifications in different screening rounds.

Columnar cell metaplasia (n = 104), cystic adenotic changes (n = 64), fibroadenomas (n = 54), and scleradenoses (n = 26) occurred most frequently among the calcifications with benign outcome (390 of 644 lesions, 60.6 %). Screening-negative high-risk lesions (59 of 644 lesions, 9.2 %) included atypical ductal hyperplasia (n = 26), lobular neoplasms (n = 13), and papilloma (n = 12). Screening-positive breast cancers (189 of 634 women, 29.8 %) resulted from DCIS diagnoses (151 of 644 lesions, 23.4 %) and invasive breast cancers (44 of 644 lesions, 6.8 %).



▶ **Fig. 1** Presentation of the study cohort. * Lack of radiological-pathological correlation (n = 6), lack of implementation of advised diagnostic excisions (n = 11), lack of implementation of recommended follow-ups after biopsy (n = 24), summarized reasons such as invasive clarification of calcifications in association with masses or architectural distortions, biopsy indication did not correspond to the recall lesion and resulted from magnification images during assessment (n = 14). # Women without breast cancer or women with breast cancer not resulting from a calcification-related lesion. DCIS: ductal carcinoma in situ. High-risk lesions: In the case of atypical epithelial proliferations of the ductal type, the final histology was based on surgical histology such as atypical ductal hyperplasia. For lesions such as flat epithelial atypia, papilloma and radial scars, an individual indication was made regarding residues of lesions and atypia.

Human reading revealed a lesion-specific PPV3 of 30.3 % (195/644), increasing from 21.2 % to 57.7 % to 100 % across categories 4a, 4b, and 5 (▶ **Table 1**). Among calcification-indicated biopsies, category 4a dominated with 76.1 % (490/644). The proportion of DCIS of high nuclear grade and invasive cancers increased across reporting categories 4a, 4b, and 5 with 5.9 % (29/490), 22.1 % (33/149), 60 % (3/5), and 3.9 % (19/490), 16.1 % (24/149), 20 % (1/5), respectively.

AI Performance

Of 195 screen-detected, calcification-associated malignancies, 14 were not detected as lesions in lesion-specific AI assessment including manual selection (score 0).

AI-positive lesions (score > 0), had a lesion-specific PPV3 of 20.8 % (91/437) in category 4a, 57.8 % (85/147) in category 4b, and 100 % (5/5) in category 5. The lesion-specific PPV3 of AI across all categories was 30.7 % (181/589) (▶ **Table 2**).

The lesion-specific false-negative AI rate was 7.2 % (95 % CI: 4.3 %, 11.4 %), corresponding to a sensitivity of 92.8 %. The non-AI-detected breast cancer cases included 13 cases of DCIS (low grade n = 3, intermediate grade n = 6, high grade n = 4) and one case of invasive breast cancer. Thirteen out of 14 were category 4a cases; the morphology (▶ **Table 2**) amorphous (amorphous n = 12 (85.7 %), granular n = 1 (7.1 %), linear n = 1 (7.1 %)) and the distribution calcification clusters (clustered n = 8 (57.1 %), segmental n = 3 (21.4 %), regional n = 2 (14.3 %), linear n = 1 (7.1 %)) were most frequent.

▶ **Table 1** Lesion-specific positive predictive value of the invasive assessment of screen-detected calcifications.

Screen-detected calcifications*	Category 4a n = 490 (100 %)	Category 4b n = 149 (100 %)	Category 5 n = 5 (100 %)	Total n = 644 (100 %)
No breast cancer	386 (78.8)	63 (42.3)	0 (0)	449 (69.7)
Benign lesions	335 (68.4)	55 (36.9)	0 (0)	390 (60.6)
Lesions with uncertain malignant potential**	51 (10.4)	8 (5.4)	0 (0)	59 (9.2)
Breast cancer (DCIS + invasive breast cancer)	104 (21.2)	86 (57.7)	5 (100)	195 (30.3)
DCIS G1	17 (3.5)	8 (5.4)	0 (0)	25 (3.9)
DCIS G2	39 (8.0)	21 (14.1)	1 (20.0)	61 (9.5)
DCIS G3	29 (5.9)	33 (22.1)	3 (60.0)	65 (10.1)
Invasive breast cancer	19 (3.9)	24 (16.1)	1 (20.0)	44 (6.8)
Lesion-specific PPV3 of human reading (%)	21.2 (104/490)	57.7 (86/149)	100.0 (5/5)	30.3 (195/644)

Unless otherwise stated, data represent absolute frequencies (percentages).
DCIS: ductal carcinoma in situ, G1: low nuclear grade, G2: intermediate nuclear grade, G3: high nuclear grade. PPV3: positive predictive value of performed invasive assessment.
* All calcification-related lesions were confirmed by vacuum-assisted biopsy. In case of a surgical intervention, the final histology was used for evaluation. For benign lesions, a biennial negative follow-up was present.
** In the case of atypical epithelial proliferation of the ductal type, final histology was based on post-surgical histology. For other lesion types, such as flat epithelial atypia, papillomas and radial scars, an individual decision was made regarding surgery depending on lesion residues and atypia.

▶ **Table 2** Lesion-specific positive predictive value of biopsy-confirmed screen-detected calcifications based on a retrospective AI evaluation.

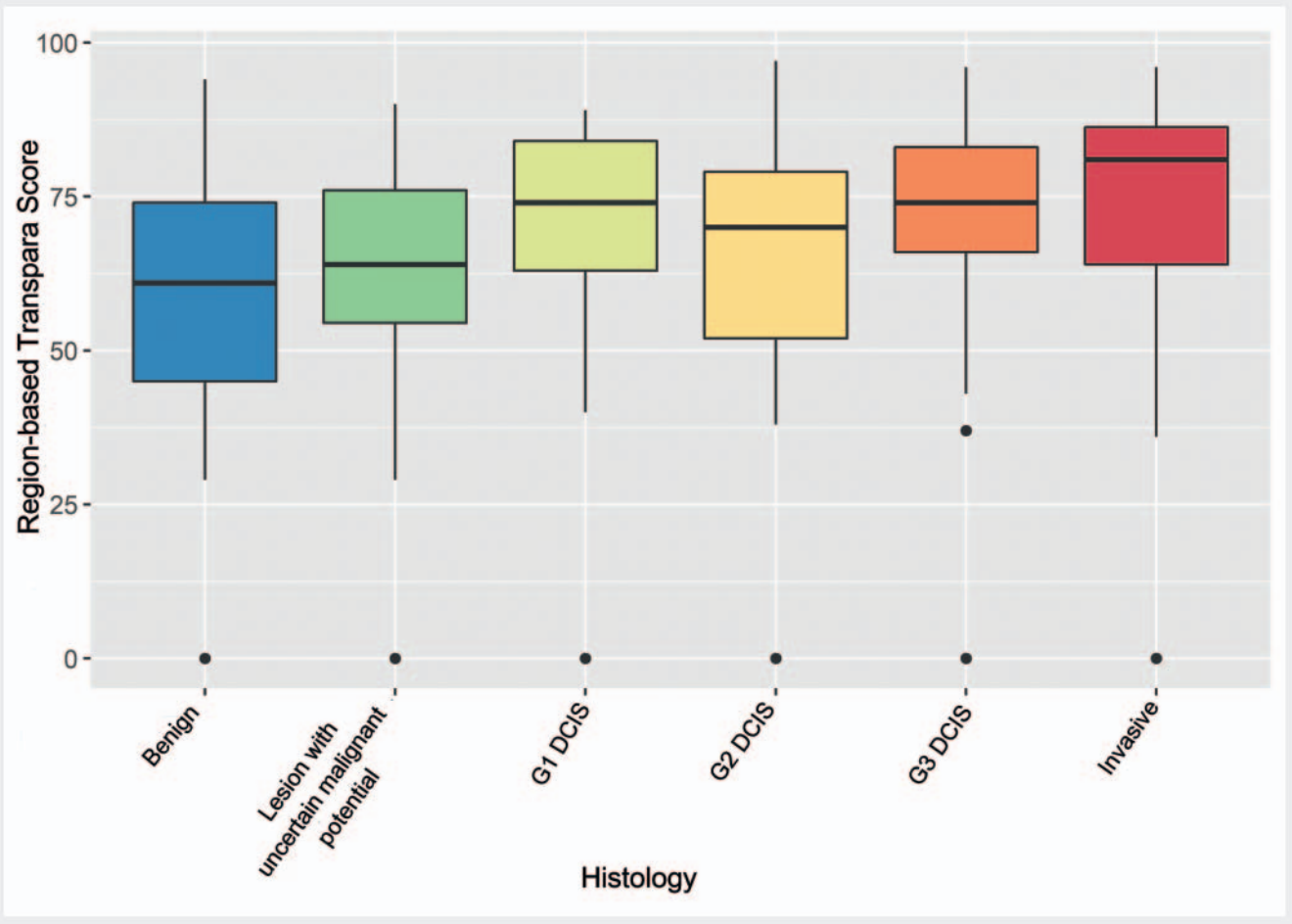
AI assessment of screen-detected calcifications	Category 4a n = 490	Category 4b n = 149	Category 5 n = 5	Total n = 644
No breast cancer	386 (100)	63 (100)	0 (0)	449 (100)
Benign lesions with region score = 0 (true-negative)	40 (10.4)	1 (1.6)	0 (0)	41 (9.1)
Benign lesions with region score > 0 (false-positive)	346 (89.6)	62 (98.4)	0 (0)	408 (90.9)
Breast cancer (DCIS + invasive breast cancer)	104 (100)	86 (100)	5 (100)	195 (100)
Malignant lesions with region score > 0 (true-positive)	91 (87.5)	85 (98.8)	5 (100)	181 (92.8)
Malignant lesions with region score = 0 (false-negative)	13 (12.5)	1 (1.2)	0 (0)	14 (7.2)
Lesion-specific PPV3 AI (%)	20.8 (91/437)	57.8 (85/147)	100 (5/5)	30.7 (181/589)

Unless otherwise stated, data represent absolute frequencies (percentages).
AI: artificial intelligence; DCIS: ductal carcinoma in situ; PPV3: positive predictive value of performed invasive assessment.

No score was shown (score = 0) for 41 of 449 calcification-associated screening-negative lesions. The rate of true-negative AI ratings was 9.1 % (95 % CI: 6.6 %, 11.9 %).

The AI-system assigned the lowest median score (61, interquartile range: 45–74) to benign lesions and the highest median score (81, 64–86) to invasive cancers. DCIS led to median scores of 74 (63–84), 70 (52–79), and 74 (66–83) with increasing nuclear grade. The distribution of the AI scores showed a strong overlap between the different histological lesions (▶ **Fig. 2**).

Only malignant lesions were recorded in the score group 96–100, which included 1.1 % (n = 7) of all 644 lesions. In the adjacent 91–95 and 86–90 score groups, the malignancy percentage decreased to 77.8 % (14 of 18) and 53.5 % (23 of 43), respectively. In the subsequent descending score groups, the malignancy percentage decreased steadily to 31.3 % (score group 65–70 (21 of 67)). In score groups 61–65 to 26–30, the malignancy percentage varied from 0 % to 25 %. There were 25.5 % (14 of 55) malignant lesions in the score 0 group. Calcification-associated invasive cancers were distributed with varying proportions across 13 of 16 score groups (▶ **Fig. 3**).



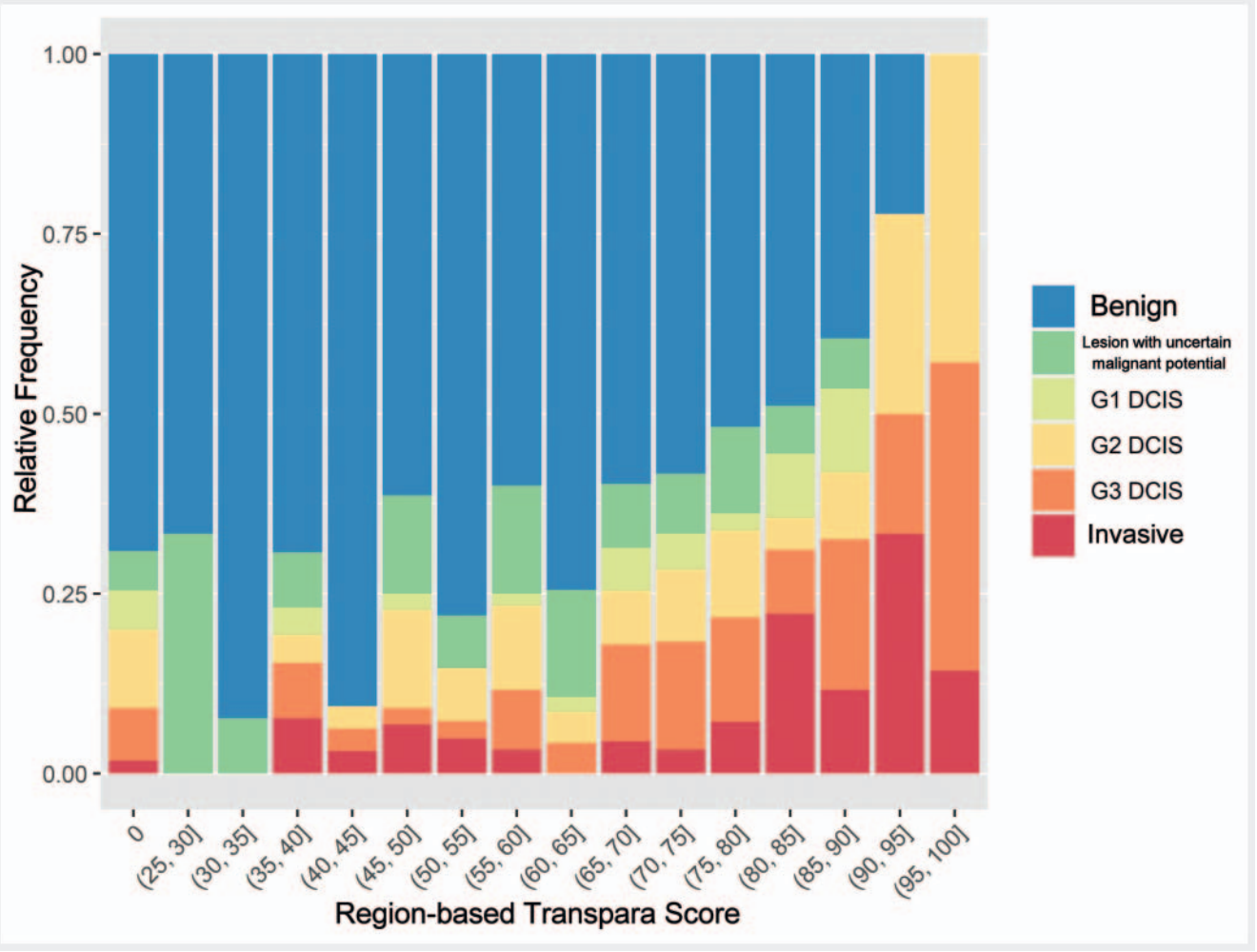
▶ **Fig. 2** Region-based AI scores of biopsy-confirmed calcifications based on digital screening mammograms in relation to the final histology. Lesions with uncertain malignant potential: In the case atypical epithelial proliferation of the ductal type, the final histology was based on post-surgical histology. For other lesions such as flat epithelial atypia, papillomas and radial scars, an individual recommendation was made with regard to surgery depending on lesion residues and atypia. DCIS: ductal carcinoma in situ, G1: low nuclear grade, G2: intermediate nuclear grade, G3: high nuclear grade.

Discussion

Clinical studies have demonstrated that AI increases radiologists' sensitivity and potentially specificity when evaluating mammograms [5]. Validation studies are needed in order to assess AI performance in different diagnostic processes [4]. The present study tested the diagnostic value of an AI application towards histologically assessed calcifications. In contrast to other validation studies, the assessment of the AI system was performed at the lesion level rather than at the mammogram level [5, 10, 12]. Therefore, the performance of the AI-system was evaluated based on preselected, specific regions. The present work on the positive predictive value of performed invasive assessment of calcifications (PPV3) complements AI validation regarding the positive predictive value of recall (PPV1) [11]. With lower malignancy rates in calcifications than in masses, there is a need for AI-applications that enable performance increase and allow for reduction of unnecessary invasive assessments of benign lesions [14].

In the present study, the PPV3 for calcifications based on human assessment was 30 % and was comparable to the AI system (31 %) when set to the lowest threshold with a false-negative rate of 7 % and a true-negative rate of 9 %. Detailed examination of the AI assessment showed increasing PPV3 values with increasing categories (4a: 21 %, 4b: 58 %, 5: 100 %), consistent with the human assessment and the current literature [8].

There are only a few validation studies evaluating AI regarding its positive prediction towards calcification based on histological assessment. Using another AI system, the probability of malignancy was retrospectively visually categorized by radiologists and lesion-specifically compared with the AI system at a 10 % threshold. There was no significant difference between the area under the receiver-operator-characteristic-curve (AUC) regarding malignancy scores and categorizations between readers and AI [15]. The results are in line with previous work, stating that neural networks can achieve over 98 % accuracy in categorizing suspicious calcifications [16].



▶ **Fig. 3** Relative frequencies of lesions within available score group of region-based AI evaluation of calcifications. Lesions with uncertain malignant potential: In the case of atypical epithelial proliferations of the ductal type, the final histology was based on post-surgical histology. For other lesions, such as flat epithelial atypia, papillomas and radial scars, an individual recommendation regarding surgical procedures was made depending on lesion residues and atypia. DCIS: ductal carcinoma in situ, G1: low nuclear grade, G2: intermediate nuclear grade, G3: high nuclear grade.

Our study showed that highly suspicious calcifications were consistantly rated with a high degree of reliability by the AI system with a score >0 (category 4b: false-negative 1/86, category 5: 0/5). In contrast, there was a higher absolute and relative number of AI false-negative calcification assessments (13/104) in cases of lower suspicion of malignancy (4a). Even though the detection of highly suspicious lesions is essential, also for medico-legal reasons, screen-detected calcifications associated with malignant lesions show a higher absolute frequency in category 4a (104 of 195 calcification-associated malignant histologies). Screening efficiency and ultimately patients would strongly benefit from an AI-system that yields an increase in the PPV3 and thus allows to reduce the number of invasive assessments of calcifications depicting benign lesions [17].

Score ranges between 96–100 and 91–95 indicated a high malignancy rate of 100 % and 77.8 %, respectively. In contrast, our validation indicated that a single score value ≤ 90 was less specific with varying malignancy percentages from 0 % to 54 %. A specific score threshold indicating the probability of malignancy could not be derived in the dominant proportion of all suspicious calcifications. The histological complexity of calcification-associated lesions might be the reason [7, 17]. Clustered amorphous calcifications of category 4a are most common among invasive assessments of calcifications [15].

A differentiation of calcifications with regard to grading of DCIS or invasive breast cancers based on the AI score was not possible with the used version of the AI system. It would be beneficial if AI could reliably distinguish delete detect in particular DCIS of intermediate and high degree as well as invasive breast cancer [18]. Among the AI false-negative lesions DCIS were most prominent (92.9 %, 13/14), regardless of grade. Further, there was one invasive breast cancer among the AI false-negatives (7.1 %, 1/14). Prospective studies of AI use in mammography-based screening are needed in order to evaluate the potential of AI to optimize the rate of unnecessary assessments while increasing the rate of biologically relevant diagnoses with regard to breast cancer mortality [9, 19].

The particular strength of the present work is the lesion-based AI assessment on a high number of cases. The diagnostic assessment as well as histological findings were subject to a high degree of standardization with follow-up. The study data were not part of the data set used to train the AI system.

A limitation is that the study design was not structured to test AI-related detection of additional, calcification-associated, malignant lesions in addition to the biopsied lesions, or to increase the sensitivity for calcification-associated malignancies, since there was no comparison with interval cancers available. The use of AI to reduce invasive assessments of benign calcifications requires further studies including interval cancers [20]. One retrospective study demonstrated that up to 50.9 % of interval cancers can be detected by an AI-system at the time of screening [21]. Yet, neither the amount of additional AI false-positive calcifications nor its prospective use with automated lesion display was tested. These results could only be related to our findings if initial diagnostic requirements, such as exclusion of associated masses, would be aligned.

In summary, the AI-system's PPV3 for calcifications in total and per biopsy-indicating category was comparable to the reader-dependent evaluation. The implementation of AI did not lead to a reduction in screening-negative calcifications leading to invasive assessment without breast cancer detection. The category-based evaluation of the AI performance revealed false-negative evaluations in the group with less suspicious lesions, particularly in category 4a, yet with a reduction of false-positive biopsies. An AI-score based histological differentiation could not be derived based on the present results.

CLINICAL RELEVANCE

Compared to human assessment, the implementation of AI does not lead to increases in positive predictive values for invasive assessment of calcification-related lesions across BIRADS-based categories.

Especially at the lowest radiological suspicion level, a dedicated human evaluation seems reasonable due to a potentially higher risk of an AI false-negative evaluation than in the more suspicious categories.

Funding

EU INTERREG V A programme Germany-Netherlands; project InMediValue 122 207

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] Perry N, Broeders M, de Wolf C et al. (eds). European guidelines for quality assurance in breast cancer screening and diagnosis. Luxembourg: Office for Official Publications of the European Communities; 2006

[2] Khil L, Heidrich J, Wellmann I et al. Incidence of advanced-stage breast cancer in regular participants of a mammography screening program: a prospective register-based study. BMC Cancer 2020; 20: 1–9

[3] Katalinic A, Eisemann N, Kraywinkel K et al. Breast cancer incidence and mortality before and after implementation of the German mammography screening program. Int J Cancer 2020; 147: 709–718

[4] Bennani-Baiti B, Baltzer PAT. Künstliche Intelligenz in der Mammadiagnostik. Radiologe 2020; 60: 56–63

[5] Hickman SE, Woitek R, Le EPV et al. Machine Learning for Workflow Applications in Screening Mammography: Systematic Review and Meta-Analysis. Radiology 2022; 302: 88–104

[6] Weigel S, Decker T, Korsching E et al. Calcifications in digital mammographic screening: improvement of early detection of invasive breast cancers? Radiology 2010; 255: 738–745

[7] Tse GM, Tan PH, Pang AL et al. Calcification in breast lesions: pathologists' perspective. J Clin Pathol 2008; 61: 145–151

[8] D'Orsi CJ, Mendelson EB, Ikeda DM et al. (eds). Breast Imaging Reporting and Data System: ACR BI-RADS – breast imaging atlas. Reston: American College of Radiology; 2003

[9] Jahresbericht Evaluation 2019. Deutsches Mammographie-Screening-Programm. Kooperationsgemeinschaft Mammographie, Berlin, November 2021. Im Internet: https://www.mammo-programm.de/download/downloads/berichte/neu_KOOPMAMMO_Jahresbericht_Eval_2019_20211112_web-Einzelseite_2.pdf

[10] Rodriguez-Ruiz A, Lång K, Gubern-Merida A et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. J Natl Cancer Inst 2019; 111: 916–922

[11] Kerschke L, Weigel S, Rodriguez-Ruiz A et al. Using deep learning to assist readers during the arbitration process: a lesion-based retrospective evaluation of breast cancer screening performance. Eur Radiol 2022; 32: 842–852

[12] Rodríguez-Ruiz A, Krupinski E, Mordang JJ et al. Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. Radiology 2019; 290: 305–314

[13] Weigel S, Decker T, Korsching E et al. Minimalinvasive biopsy results of "uncertain malignant potential" in digital mammography screening: high prevalence but also high predictive value for malignancy. Fortschr Röntgenstr 2011; 183: 743–748

[14] Burnside ES, Ochsner JE, Fowler KJ et al. Use of calcification descriptors in BI-RADS 4th edition to stratify risk of malignancy. Radiology 2007; 242: 388–395

[15] Do YA, Jang M, Yun B et al. Diagnostic Performance of Artificial Intelligence-Based Computer-Aided Diagnosis for Breast Microcalcification on Mammography. Diagnostics 2021; 11: 1409. doi:10.3390/diagnostics11081409

[16] Schönenberger C, Hejduk P, Ciritsis A et al. Classification of Mammographic Breast Microcalcifications Using a Deep Convolutional Neural Network: A BI-RADS-Based Approach. Invest Radiol 2021; 56: 224–231

[17] Tot T, Gere M, Hofmeyer S et al. The clinical value of detecting calcifications on a mammogram. Semin Cancer Biol 2021; 72: 165–174

[18] Maxwell AJ, Hilton B, Clements K et al. Unresected screen-detected ductal carcinoma in situ: Outcomes of 311 women in the Forget-Me-Not 2 study. Breast 2022; 61: 145–155

[19] Wallis MG. Artificial intelligence for the real world of breast screening. Eur J Radiol 2021; 144: 109661. doi:10.1016/j.ejrad.2021.109661

[20] Lang K, Hofvind S, Rodriguez-Ruiz A et al. Can artificial intelligence reduce the interval cancer rate? Eur Radiol 2021; 31: 5940–5947

[21] Wanders AJT, Mees W, Bun PAM et al. Interval cancer detection using a neural network and breast density in women with negative screening mammograms. Radiology 2022; 303: 269–75