Künstliche Intelligenz zur Indikationsstellung einer invasiven Mikrokalkabklärung im Mammografie-Screening

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

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

Institute
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

eingereicht 08.07.2022
akzeptiert 16.10.2022
online publiziert 2022

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

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

Korrespondenzadresse
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

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.

Kernaussagen:

- Der PPV3 der Mikrokalkabklärung ist unter KI-Bewertung vergleichbar zur menschlichen Bewertung.
- Die Befundstufe 4a unterliegt der ausgeprägtesten KI-induzierten Minderung Screening-positiver sowie Screening-negativer Läsionen.
- Die Score-Werte diskriminieren keine Subgruppen histologischer Läsionen.

Zitierweise

- 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

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, inter-quartile 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.

Einleitung

Die Mammografie gilt als einzige evidenzbasierte Methode zur systematischen Brustkrebs-Früherkennung mit nachgewiesenem Effekt auf die Senkung der Brustkrebs-spezifischen Sterblichkeit und ist in Deutschland flächendeckend etabliert sowie wissenschaftlich belegt wirksam [1–3].

Künstliche Intelligenz (KI) verwendet unterschiedliche Algorithmen für die Lösung verschiedener Aufgabenstellungen und kann Menschen Entlastung oder Unterstützung bieten [4]. Die Weiterentwicklung von Computer-Aided-Detection (CAD)-Systemen infolge technischer Fortschritte und Deep-lerning-Algorithmen kann die Leistungsfähigkeit in der mammografischen Früherkennung steigern. Eine Metaanalyse retrospektiver Triage-Studien zeigt, dass die alleinige Applikation von KI-Algorithmen die Anzahl von Befundern bewerteten Mammografien zwischen 17 %–91 % senken kann, während die Minderung der Brustkrebsdetektion 0 %–7 % beträgt [5].

In der Altersgruppe 50–69 Jahre bilden Mikroverkalkungen die zweithäufigste mammografische Auffälligkeit, die zur weiteren Abklärungsdiagnostik führt und zugleich die zweithäufigste mammografische Auffälligkeit in der Detektion von Brustkrebs [6, 7]. Mikroverkalkungen repräsentieren ein breites Läsionsspektrum: von mastopathischen Mammaläsionen über Risikoläsionen und Vorläuferläsionen des invasiven Brustkrebses bis zum invasiven Mammakarzinom mit variierender biologischer Bedeutung und variierendem positiven prädiktiven Wert in der invasiven Abklärungsdiagnostik (PPV3) [8, 9].

Aus Anwendersicht sind Überprüfungen der diagnostischen Wertigkeit eines individuellen KI-Algorithmus in der Dignitätsbeurteilung von Mikroverkalkungen bedeutsam, um die gewonnene abstrakte KI-Information zu einer definierten mammografischen Läsion validiert in den finalen menschlichen Entscheidungsprozess integrieren zu können. Die retrospektive Integration eines verfügbaren KI-Systems [10] in den Entscheidungsprozess der Konsensuskonferenz hatte das Potenzial falsch positive Rückrufe zur Abklärungsdiagnostik zu mindern, allerdings zeigten Mikrokalk-assoziierte Läsionen eine geringere Sensitivität als Herd-assoziierte Läsionen [11].

Ziel der vorliegenden retrospektiven Studie war es, die diagnostische Wertigkeit eines individuellen KI-Algorithmus in der Bewertung des Malignitätsverdachtsgrades Screening-detektierter, histologisch abgeklärter Mikroverkalkungen in Bezug zur menschlichen Befundung zu prüfen.

Material und Methoden

Die Studie umfasste 634 Frauen mit Mikrokalk-bedingter invasiver Abklärung nach Teilnahme in einer Mammografie-Screening-Einheit von Juli 2012 bis Juni 2018. Die histologisch abgeklärten Mikrokalkbefunde wurden retrospektiv einer gezielten KI-Bewertung auf Läsionsebene unterzogen. Diese wurde hinsichtlich des radiologischen Verdachtsgrades für Malignität evaluiert und anhand des Goldstandards, der finalen Histologie, kategorisiert.

Die Arbeit wurde im Rahmen des EU Projektes INTERREG V A, InMediValue 122207 durchgeführt. Es wurde ein Votum der Ethikkommission der Ärztekammer Westfalen-Lippe und der Westfälischen Wilhelms-Universität eingeholt, die keine Bedenken ethischer oder rechtlicher Art gegen die Durchführung des Forschungsvorhabens hatte.

Screening-Prozess

Im Rahmen des deutschen Mammografie-Screening-Programms werden Frauen zwischen 50 und 69 Jahren per Brief zu einer digitalen Mammografie-Screening-Untersuchung in 2 Ebenen eingeladen. Die Screening-Mammografien werden durch 2 zertifizierte Befunder räumlich und zeitlich unabhängig voneinander ausgewertet. Bei mindestens einer Auffälligkeit diskutieren beide Befunder den Fall in einer Konsensuskonferenz gemeinsam mit dem sogenannten programmverantwortlichen Arzt. Dieser entscheidet abschließend, ob ein Rückruf zur weiteren Abklärungs-

diagnostik indiziert ist und führt die folgende Diagnostik mit Indikationsstellung zur invasiven Abklärung durch [1].

Klinische Studiendaten

Für die Studie wurde zur Graduierung der Malignitätswahrscheinlichkeit die in der Konsensuskonferenz dokumentierte Befundstufe (4a, 4b, 5) aus der Screening-Software MaSc verwendet (KVWL, Dortmund, Germany). Diese orientierte sich an dem Breast Imaging Reporting and Data System (BI-RADS) Version 4 [8].

Die Erstellung der Mammografie in 2 Ebenen erfolgte an 2 Standorten (Sectra MDM L30, Linköping, Schweden; Philips MDM L50, Philips Healthcare, Einhoven, Niederlande; Hologic 3Dimensions, Marlborough, MA, US; Mammomat Inspiration, Mammomat Revelation, Siemens Healthcare, Erlangen, Germany). Die unabhängige Doppelbefundung wurde von 5 Befundern, inklusive 2 programmverantwortlichen Ärzten, umgesetzt. Die standardisierte Abklärungsdiagnostik der Mikrokalk-assoziierten Läsionen umfasste die Sonografie zum Ausschluss assoziierter Herdläsionen (Acuson S2000, Siemens Healthcare, Erlangen, Germany) und Vergrößerungsaufnahmen in cranio-caudaler und lateraler Projektion (Hologic Selenia Dimensions, Marlborough, MA, US). Für suspekte, reine Mikroverkalkungen wurde als Methode der ersten Wahl eine röntgengesteuerte Vakuumbiopsie (Hologic Multicare Platinum, Marlborough, MA, US) geplant.

Datenerhebung

Verwendet wurde die CE und FDA-zertifizierte KI-basierte Software Transpara (Version 1.7.0) der Firma ScreenPoint Medical, Netherlands. Eingesetzt wurde ein Deep-learning-Algorithmus, welcher auf einem tiefen neuronalen Faltungsnetzwerk (deep convolutional neural network) basiert. Der Algorithmus wurde anstatt auf Bilddaten von über 2 Millionen histologisch bestätigten Läsionen trainiert und externen, klinischen Validierungen unterzogen [10]. Anhand der schriftlichen und bildlichen Dokumentation wurden die histologisch abgeklärten Mikroverkalkungen in der Screening-Mammografie durch einen programmverantwortlichen Arzt reidentifiziert. Die Mikrokalkmorphologie und die Mikrokalkanordnung wurden bestimmt [8]. Erhoben wurde der läsionen-spezifische KI-Score zwischen 1 und 100 durch eine anwenderbezogene Anwahl per Mausklick, falls nicht automatisiert angezeigt. 100 repräsentiert den höchsten Auffälligkeitsgrad für Malignität [12]. Für Läsionen, die von dem System mit einem Score ≤ 28 bewertet wurden, wird dem Nutzer kein Score angezeigt (analysiert als Score = 0). Für Läsionen, die mit 98–100 bewertet wurden, wird dem Nutzer ein Score von 98 angezeigt. Im Fall variierender Scores einer Läsion in den unterschiedlichen mammografischen Projektionsebenen wurde der höhere Score verwendet.

Einschlusskriterien, Ausschlusskriterien

Inkludiert wurden mittels digitaler Vollfeldmammografie-Technik im Screening detektierte, Mikrokalk-assoziierte Läsionen mit finaler Histologie, für die eine Vakuumbiopsie mit bestehender positiver radiologisch-pathologischer Konkordanzprüfung durchgeführt wurde. Falls nach interdisziplinärer Diskussion angezeigt, wurden weitere invasive Maßnahmen geplant. Für einen Studieneinschluss mussten alle Empfehlungen abgeschlossen sein. Ohne Brustkrebsnachweis musste ein zweijähriges mammografisches Follow-up vorliegen. Die resultierenden Ausschlusskriterien sind in ▶ **Abb. 1** benannt.

Screening-positive Mikrokalkläsionen

Zu den Screening-positiven Läsionen zählten das duktale Carcinoma in situ (DCIS) und das invasive Mammakarzinom. Verwendet wurde das finale postoperative Ergebnis. Bei neoadjuvanter Therapie wurde das Ergebnis der minimalinvasiven Beurteilung herangezogen. Brustkrebsfälle wurden nach dem Kernmalignitätsgrad oder dem Grading differenziert.

Screening-negative Mikrokalkläsionen

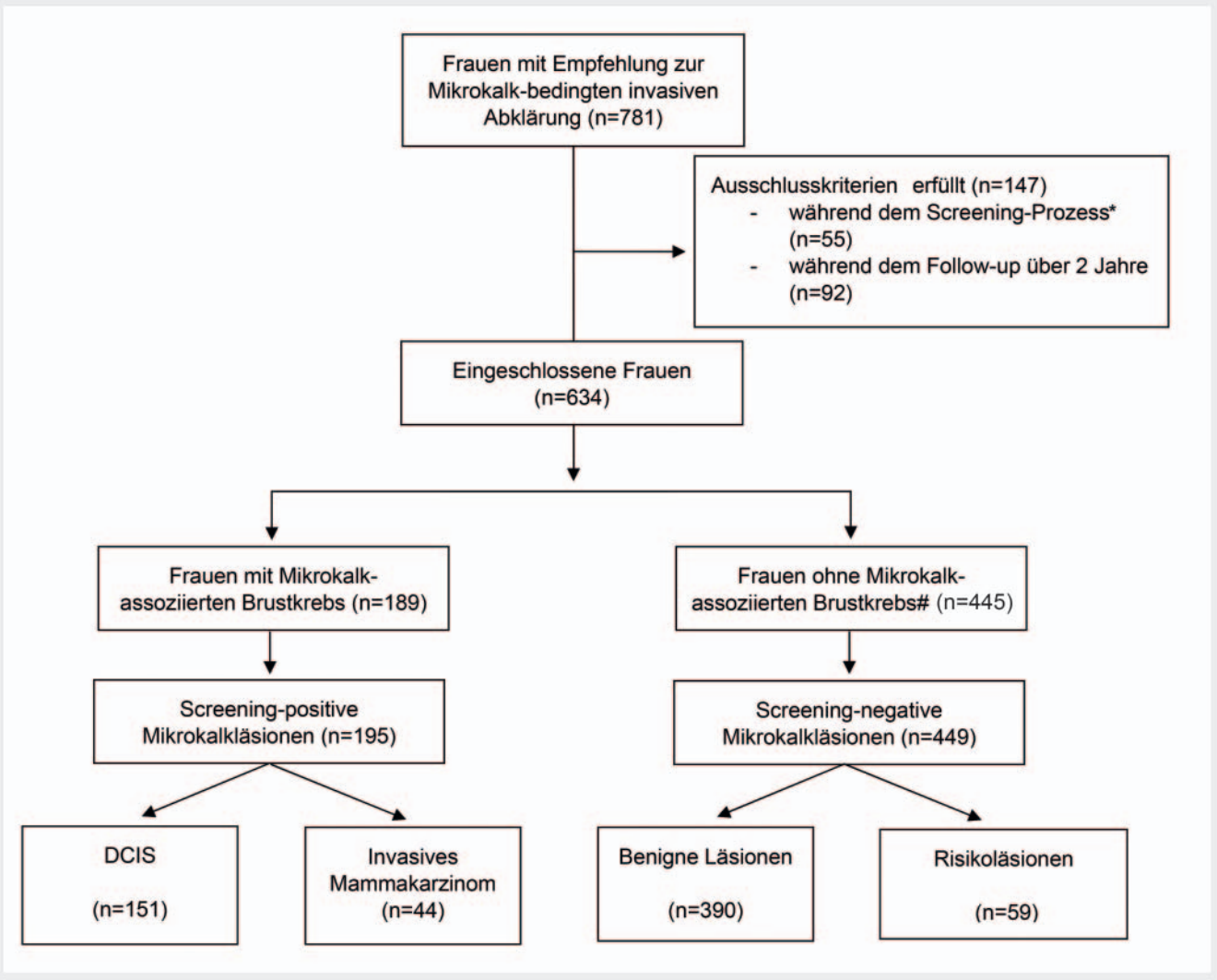
Zu den Screening-negativen Läsionen zählten histologisch benigne Läsionen. Konform zur Screening-Evaluation wurden zudem Läsionen unklaren malignen Potenzials (Risikoläsionen) Screening-negativ gewertet. Wenn eine zusätzliche diagnostische Exzision indiziert wurde, wurde das postoperative abschließende, histologische Ergebnis verwendet. Indikationen einer diagnostischen Exzision lagen in jedem Fall von atypischen Epithelproliferationen vom duktalen Typ vor sowie bei residualen Läsionsanteilen einer flachen epithelialen Atypie (FEA), bei Papillomen und radiären Narben [13]. Screening-negative Bewertungen basierten zudem auf einem zweijährigen negativen Follow-up durch eine weitere mammografische Diagnostik ohne Brustkrebsdetektion.

KI-negative und KI-positive Mikrokalkläsionen

KI-negativ zählten Läsionen, die in der gezielten Bewertung keinen Score lieferten (Score = 0). Alle Läsionen mit einem erhebbaren Regionen-spezifischen Score (Score ≥ 29) wurden KI-positiv gewertet.

Statistische Auswertung

Die Analysen wurden mit der Statistiksoftware R (Version 4.0.2) durchgeführt. Kategoriale Parameter wurden als absolute und relative Häufigkeiten dargestellt und stetige Parameter als Median und Interquartilsabstand. Läsionen-spezifisch wurde der positive prädiktive Wert der umgesetzten invasiven Abklärungsdiagnostik (PPV3) für die befunderbezogene und die KI-bezogene Bewertung von Mikroverkalkungen ermittelt. Zur Bewertung der KI-Performance wurde die Läsionen-spezifische Rate falsch negativer Bewertungen des Systems, d. h. der Anteil KI-negativer Läsionen unter den Screening-positiven Mikrokalkläsionen (1 – Sensitivität), und die Rate KI-richtig-negativer Bewertungen, d. h. der Anteil KI-negativer Läsionen unter den Screening-negativen Mikrokalkläsionen (Spezifität), ermittelt. Für die Performance-Indikatoren wurde ein 95 %-Konfidenzintervall mittels nicht parametrischem Bootstrap berechnet.



▶ **Abb. 1** Darstellung des Studienkollektivs. * Fehlende radiologisch-pathologische Korrelation (n = 6), fehlende Umsetzung angeratener diagnostischer Exzisionen (n = 11), fehlende Umsetzung empfohlener Kontrollen nach Biopsie (n = 24), zusammengefasste Gründe wie invasive Abklärung von Mikrokalk in Assoziation zu Herden oder Architekturstörungen, die Biopsieindikation entsprach nicht der Rückrufläsion und ergab sich aus Vergrößerungsaufnahmen der Abklärungsdiagnostik (n = 14). #Frauen ohne Brustkrebs oder Frauen mit einem Mammakarzinom, das nicht aus einer kalktragenden Läsion hervorging. DCIS: duktales Carcinoma in situ. Risikoläsionen: Die finale Histologie basierte im Falle atypischer Proliferationen vom duktalen Typ in jedem Fall auf der postoperativen Histologie wie z. B. einer atypischen duktalen Hyperplasie. Bei Läsionen wie der flachen epithelialen Atypie, Papillomen und radiären Narben wurde eine individuelle Indikationsstellung bezüglich einer Operation in Abhängigkeit von Läsionsresten und Atypien getroffen.

Ergebnisse

Screeningresultate

Eingeschlossen wurden die histologischen Ergebnisse von 634 Frauen mit 644 Mikrokalk-tragenden abgeklärten Läsionen (▶ **Abb. 1**). 2 Frauen erhielten invasive Mikrokalkabklärungen in verschiedenen Screeningrunden.

Unter den Mikrokalkläsionen mit benignem Ergebnis (390 von 644 Läsionen, 60,6 %) traten am häufigsten die Kolumnarzellmetaplasie (n = 104), zystisch-adenotische Veränderungen (n = 64), Fibroadenome (n = 54) und Skleradenosen (n = 26) auf. Screening-negative Risikoläsionen (59 von 644 Läsionen, 9,2 %) umfassten die atypische duktale Hyperplasie (n = 26), lobuläre Neoplasien (n = 13) und Papillome (n = 12). Die Screening-positiven Brustkrebsfälle (189 von 634 Frauen, 29,8 %) resultierten aus DCIS-Diagnosen (151 von 644 Läsionen, 23,4 %) und invasiven Mammakarzinomen (44 von 644 Läsionen, 6,8 %).

Über alle Befundstufen betrug der läsionen-spezifische PPV3 nach menschlicher Befundung 30,3 % (195/644), er stieg über die im Screening erhobenen Befundstufen 4a, 4b und 5 von 21,2 % über 57,7 % auf 100 % an (▶ **Tab. 1**). Unter den Mikrokalk-indizierten Biopsien dominierte die Befundstufe 4a mit 76,1 % (490/644). Der Anteil des DCIS vom hohen Kernmalignitätsgrad und invasiver Karzinome nahm über die Befundstufen 4a, 4b und 5 mit 5,9 % (29/490), 22,1 % (33/149), 60 % (3/5) bzw. 3,9 % (19/490), 16,1 % (24/149), 20 % (1/5) zu.

▶ **Tab. 1** Läsionen-spezifischer positiver prädiktiver Wert der invasiven Abklärungsdiagnostik von Screening-detektierten Mikroverkalkungen.

Screening-detektierte Mikrokalklsionen*	Befundstufe 4a n = 490 (100 %)	Befundstufe 4b n = 149 (100 %)	Befundstufe 5 n = 5 (100 %)	Summe n = 644 (100 %)
Kein Brustkrebs	386 (78,8)	63 (42,3)	0 (0)	449 (69,7)
Benigne Läsionen	335 (68,4)	55 (36,9)	0 (0)	390 (60,6)
Läsionen unklaren malignes Potenzials**	51 (10,4)	8 (5,4)	0 (0)	59 (9,2)
Brustkrebs (DCIS+invasives Mammakarzinom)	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)
Invasives Karzinom	19 (3,9)	24 (16,1)	1 (20,0)	44 (6,8)
Läsionen-spezifischer PPV3 Befunder (%)	21,2 (104/490)	57,7 (86/149)	100,0 (5/5)	30,3 (195/644)

Sofern nicht anders angezeigt, stellen die Angaben absolute Häufigkeiten (Prozentwerte) dar.
DCIS: duktales Carcinoma in situ, G1: geringer Kernmalignitätsgrad, G2: intermediärer Kernmalignitätsgrad, G3: hoher Kernmalignitätsgrad; PPV3: positiver prädiktiver Wert der invasiven, umgesetzten Abklärungsdiagnostik.

* Alle Mikrokalkläsionen wurden einer invasiven Abklärung mittels Vakuumbiopsie unterzogen, eine radiologisch-pathologische Korrelation lag vor. Im Falle einer Operationsempfehlung wurde die finale Histologie gewertet. Für benigne Läsionen folgte ein zweijähriges Follow-up ohne Brustkrebsdiagnose.

** Die finale Histologie basierte für atypische Proliferationen vom duktalen Typ in jedem Fall auf der postoperativen Histologie wie z. B. einer atypischen duktalen Hyperplasie. Bei Läsionen wie der flachen epithelialen Atypie, Papillomen und radiären Narben wurde eine individuelle Indikationsstellung bezüglich einer Operation in Abhängigkeit von Läsionsresten und Atypien getroffen.

▶ **Tab. 2** Läsionen-spezifischer positiver prädiktiver Wert der invasiven Abklärungsdiagnostik von Screening-detektierten Mikroverkalkungen basierend auf einer retrospektiven KI-Bewertung.

KI-Bewertung Screening-detektierter Mikrokalkläsionen	Befundstufe 4a n = 490	Befundstufe 4b n = 149	Befundstufe 5 n = 5	Summe n = 644
Kein Brustkrebs	386 (100)	63 (100)	0 (0)	449 (100)
Benigne Läsionen mit Regionen-Score = 0 (richtig-negativ)	40 (10,4)	1 (1,6)	0 (0)	41 (9,1)
Benigne Läsionen mit Regionen-Score > 0 (falsch positiv)	346 (89,6)	62 (98,4)	0 (0)	408 (90,9)
Brustkrebs (DCIS+invasives Mammakarzinom)	104 (100)	86 (100)	5 (100)	195 (100)
Maligne Läsionen mit Regionen-Score > 0 (richtig-positiv)	91 (87,5)	85 (98,8)	5 (100)	181 (92,8)
Maligne Läsionen mit Regionen-Score = 0 (falsch negativ)	13 (12,5)	1 (1,2)	0 (0)	14 (7,2)
Läsionen-spezifischer PPV3 KI (%)	20,8 (91/437)	57,8 (85/147)	100 (5/5)	30,7 (181/589)

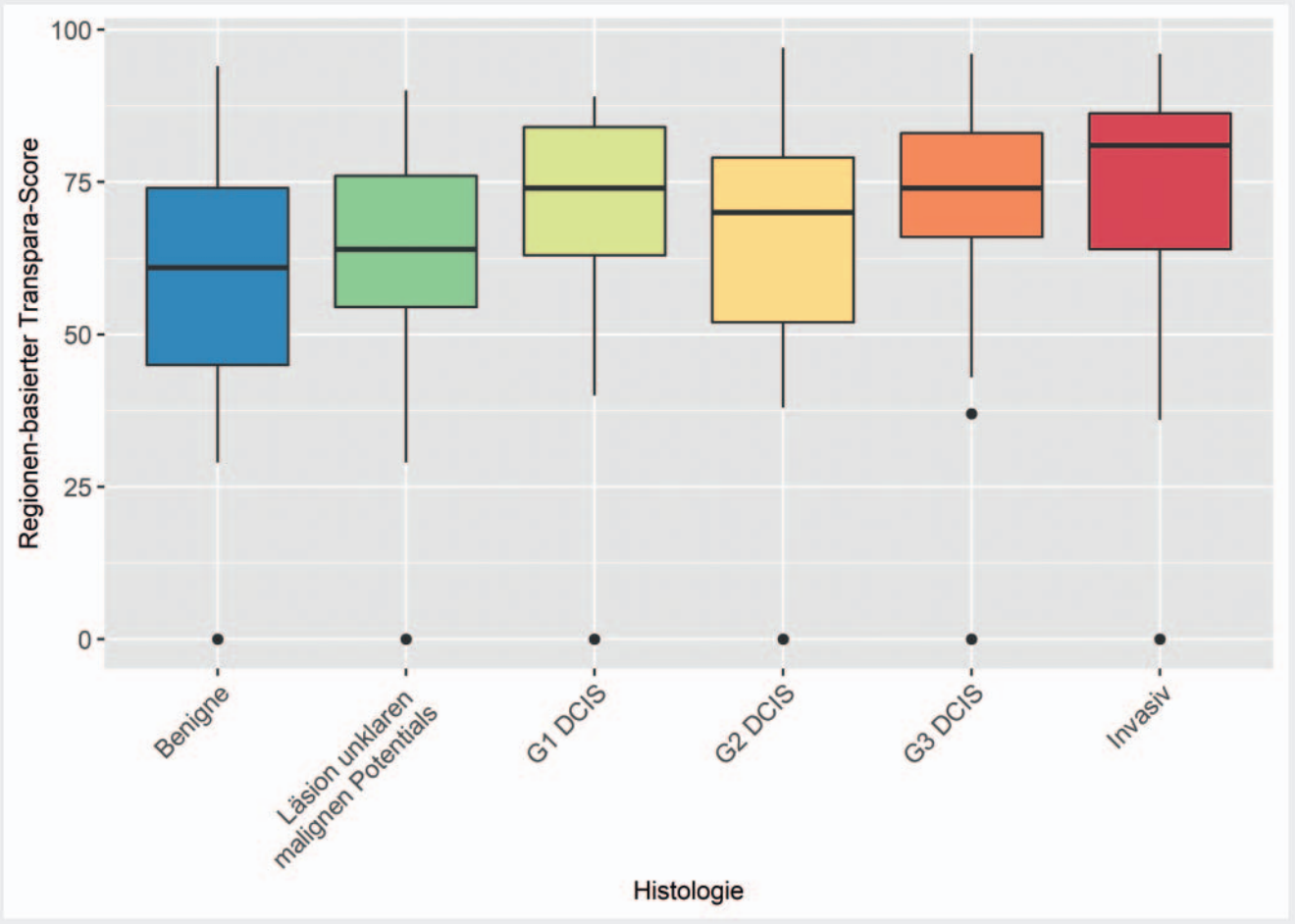
Sofern nicht anders angezeigt, stellen die Angaben absolute Häufigkeiten (Prozentwerte) dar.
KI: künstliche Intelligenz; DCIS: duktales Carcinoma in situ; PPV3: positiver prädiktiver Wert der invasiven, umgesetzten Abklärungsdiagnostik.

KI-Performance

14 von 195 Screening-detektierten, Mikrokalk-assoziierten Malignomen wurden in der läsionen-spezifischen KI-Bewertung inklusive manueller Anwahl nicht als Läsion erkannt (Score 0).

Auf Basis der KI-positiven Läsionen (Score > 0) ergab sich eine vergleichbare Verteilung des läsionen-spezifischen PPV3 pro Befundstufe zur ausschließlich menschlichen Bewertung von 20,8 % (91/437) in der Kategorie 4a, 57,8 % (85/147) in der Kategorie 4b und 100 % (5/5) in der Kategorie 5. Der läsionen-spezifische PPV3 aller Befundstufen betrug mit KI 30,7 % (181/589) (▶ **Tab. 2**).

Die läsionen-spezifische Rate falsch negativer KI-Bewertung betrug 7,2 % (95 %-CI: 4,3 %, 11,4 %), was einer Sensitivität von 92,8 % entsprach. Die nicht KI-erkannten Brustkrebsfälle umfassten 13-mal das DCIS (niedriger Kernmalignitätgrad n = 3, intermediärer Kernmalignitätsgrad n = 6, hoher Kernmalignitätsgrad n = 4) und 1-mal ein invasives Mammakarzinom, es dominierte die Befundstufe 4a (▶ **Tab. 2**), die Morphologie amorph (amorph n = 12 [85,7 %], granulär n = 1 [7,1 %], linear n = 1 [7,1 %]) und die Anordnung gruppiert (gruppiert n = 8 [57,1 %], segmental n = 3 [21,4 %], regional n = 2 [14,3 %], linear n = 1 [7,1 %]).



▶ **Abb. 2** Regionenbasierte KI-Scores der invasiv abgeklärten Mikrokalkareale basierend auf der digitalen Screening-Mammografie in Bezug zur finalen Histologie. Läsionen unklaren malignen Potenzials: Die finale Histologie basierte im Falle atypischer Proliferationen vom duktalen Typ in jedem Fall auf der postoperativen Histologie wie z. B. einer atypischen duktalen Hyperplasie. Bei Läsionen wie der flachen epithelialen Atypie, Papillomen und radiären Narben wurde eine individuelle Indikationsstellung bezüglich einer Operation in Abhängigkeit von Läsionsresten und Atypien getroffen. DCIS: duktales Carcinoma in situ, G1: geringer Kernmalignitätsgrad, G2: intermediärer Kernmalignitätsgrad, G3: hoher Kernmalignitätsgrad.

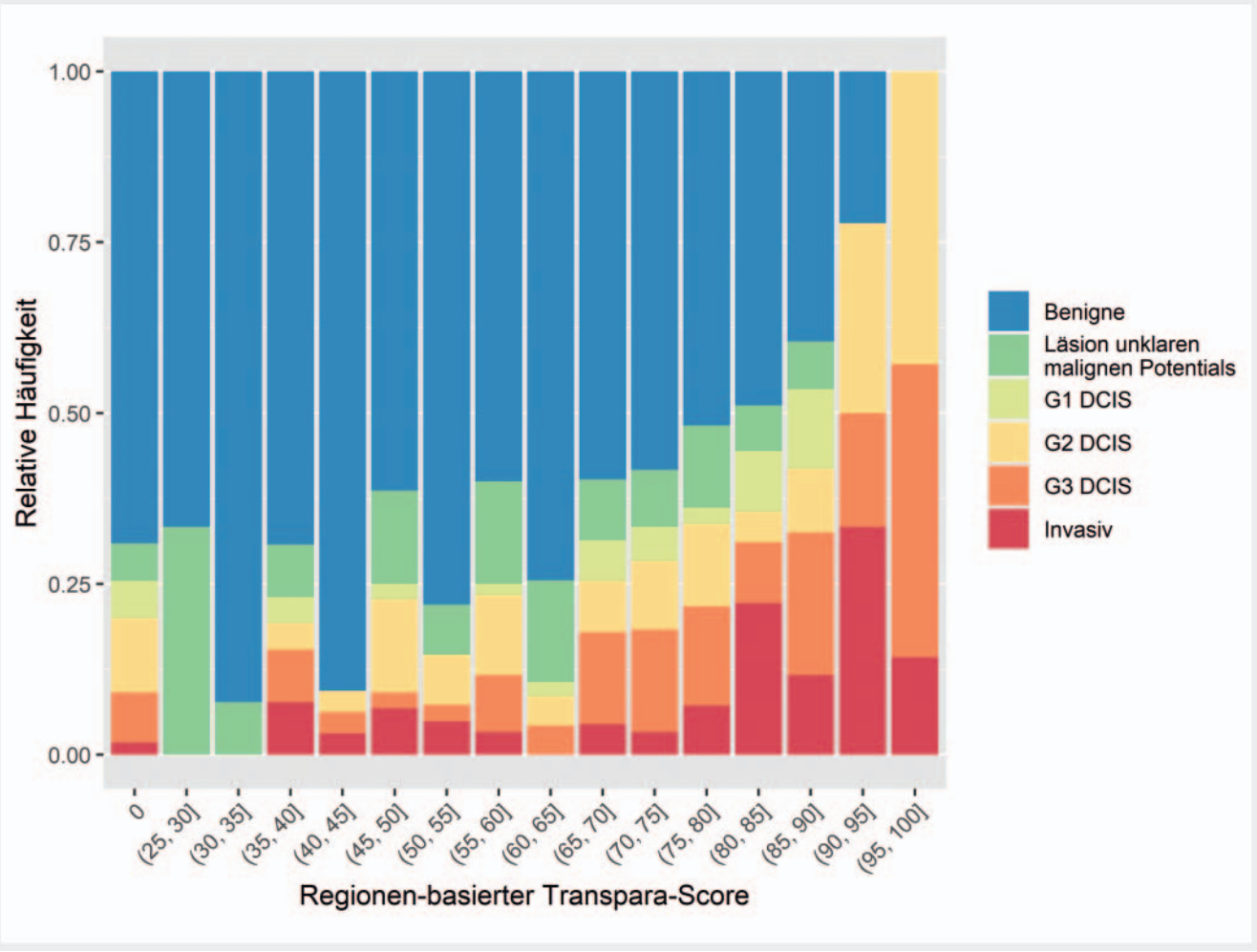
Für 41 von 449 Mikrokalk-assoziierten, Screening-negativen Läsionen wurde kein Score angezeigt (Score = 0). Die Rate richtig negativer KI-Bewertungen lag bei 9,1 % (95 %-CI: 6,6 %, 11,9 %).

Die Score-Anzeige wies für die Gruppe der benignen Läsionen den geringsten medianen Score (61, Interquartilsabstand: 45–74) und für invasive Karzinome den höchsten medianen Score (81, 64–86) auf: Das DCIS erzeugte mit steigendem Kernmalignitätsgrad mediane Scores von 74 (63–84), 70 (52–79) und 74 (66–83). Die Scores zeigten eine deutliche Überlappung der Verteilung zwischen den unterschiedlichen histologischen Läsionen (▶ **Abb. 2**).

In der Score-Gruppe 96–100, die 1,1 % (n = 7) aller 644 Läsionen inkludierte, wurden ausschließlich maligne Läsionen erfasst. In den angrenzenden Score-Gruppen 91–95 und 86–90 sank der Malignitätsanteil auf 77,8 % (14 von 18) bzw. 53,5 % (23 von 43). In den folgenden absteigenden Score-Gruppen sank der Malignitätsanteil bis zur Score-Gruppe 65–70 kontinuierlich auf 31,3 % (21 von 67). In den Score-Gruppen 61–65 bis 26–30 variierte der Malignitätsanteil zwischen 0 % und 25 %. In der Gruppe Score 0 befanden sich 25,5 % (14 von 55) maligne Läsionen. Mikrokalk-assoziierte invasive Karzinome verteilten sich mit variierendem Anteil auf 13 von 16 Score-Gruppen (▶ **Abb. 3**).

Diskussion

Klinische Studien haben gezeigt, dass KI die mammografische Befunder-Sensitivität erhöht und potenziell zur Verbesserung der Spezifizität beitragen kann [5]. Um die KI-Performance in unterschiedlichen diagnostischen Prozessen einschätzen zu können, ohne die Gesetzmäßigkeiten zu kennen, sind Validierungsstudien sinnvoll [4]. Die vorliegende Studie prüfte die diagnostische Wertigkeit einer KI-Anwendung in der Dignitätsbewertung von histologisch gesicherten Mikrokalkläsionen. Sie hebt sich von anderen Validierungen ab, da eine gezielte Bewertung des KI-Systems auf der Läsionsebene und nicht – wie in vielen Studien üblich – auf der Mammogramm-Ebene durchgeführt wurde [5, 10, 12]. Das heißt, die Performance des Systems wurde auf Basis präselektierter, spezifischer Regionen evaluiert. Die vorliegende Arbeit zum positiven prädiktiven Wert der invasiven Mikrokalkabklärung (PPV3) er-



▶ **Abb. 3** Relative Häufigkeiten von histologischen Läsionen pro verfügbarer Score-Gruppe der regionenbasierten KI-bezogenen Mikrokalkbewertung. Läsionen unklaren malignen Potenzials: Die finale Histologie basierte im Falle atypischer Proliferationen vom duktalen Typ in jedem Fall auf der postoperativen Histologie wie z. B. einer atypischen duktalen Hyperplasie. Bei Läsionen wie der flachen epithelialen Atypie, Papillomen und radiären Narben wurde eine individuelle Indikationsstellung bezüglich einer Operation in Abhängigkeit von Läsionsresten und Atypien getroffen. DCIS: duktales Carcinoma in situ, G1: geringer Kernmalignitätsgrad, G2: intermediärer Kernmalignitätsgrad, G3: hoher Kernmalignitätsgrad

gänzt die KI-Validierung bezüglich des positiven prädiktiven Wertes für den Rückruf zur Abklärungsdiagnostik (PPV1) [11]. Bei niedrigeren Malignitätsraten in der Mikrokalkabklärung als in der Herdabklärung wäre eine Steigerung durch KI-Anwendungen wünschenswert und durch Einsparung benigner Abklärungen von Relevanz [14].

In der vorliegenden Studie betrug der PPV3 für Mikroverkalkungen im inkludierten Läsionskollektiv mittels menschlicher Bewertung 30 % und lag bei niedrigster Schwelle im gezielten KI-Einsatz vergleichbar (31 %). Die Rate falsch negativer KI-Bewertungen betrug 7 %, die Rate richtig negativer KI-Bewertungen 9 %. Die Detailbetrachtung der KI-Bewertung zeigte mit ansteigender Befundstufe ansteigende PPV3-Werte von 21 % (Befundstufe 4a), 58 % (Befundstufe 4b) und 100 % (Befundstufe 5), konform zur menschlichen Bewertung der inkludierten Screening-Untersuchungen und der Literatur [8].

Gezielte Validierungsstudien zu umgesetzten histologischen Abklärungen von Mikrokalk sind selten. Unter Verwendung eines anderen KI-CAD-Systems wurde retrospektiv durch Radiologen die Wahrscheinlichkeit für Malignität visuell kategorisiert und läsionen-spezifisch mit einem Cutt-off von 10 % mit der KI-Bewertung verglichen. Die Studie fand keinen signifikanten Unterschied zwischen den AUC- (area under the receiver operator characteristic curve) Werten bezüglich Malignitäts-Scores und Kategorisierungen zwischen Befundern und KI [15]. Die Ergebnisse stehen im Einklang damit, dass neuronale Netzwerke eine Genauigkeit in der suspekten Mikrokalkkategorisierung von über 98 % erreichen können [16].

Konform zeigte unsere Studie, dass höhergradig und hochgradig suspekte Mikrokalkläsionen mit hoher Zuverlässigkeit durch die KI mit einem Score > 0 bewertet wurden (Befundstufe 4b: falsch negativ 1/86, Befundstufe 5: 0/5). Dagegen ergab sich bei geringergradigem Malignitätsverdacht (4a) eine höhere absolute und relative Anzahl KI-falsch negativer Mikrokalkbewertungen (13/104). Auch wenn das Erkennen höhergradig suspekter Läsionen u. a. aus medico-legalen Gründen essenziell ist, weisen mali-

gne Mikrokalkläsionen im Screening ein häufigeres absolutes Vorliegen in der Befundstufe 4a auf (104 von 195 Mikrokalk-assoziierten malignen Histologien). Wünschenswert wäre durch KI eine Reduktion benigner Mikrokalkabklärungen mit Anhebung des PPV3 insbesondere in dieser Kategorie [17].

Score-Werte zwischen 96–100 und 91–95 wiesen eine hohe Malignitätsrate von 100 % bzw. 77,8 % auf. Dagegen war die Einordnung eines einzelnen Score-Wertes ≤ 90 in unserer Validierung uneindeutiger bei variierenden Malignitätsanteilen von 0 % bis 54 %. Ein Schwellenwert oder eine eindeutige Graduierung für Malignität ließ sich im dominierenden Anteil aller auffälliger Mikrokalkabklärungen nicht ableiten. Die histologische Komplexität der Mikrokalk-assoziierten Läsionen mit einer teils mammografisch unspezifischen Mikrokalkproduktion mag ursächlich sein [7, 17]. Amorphe, gruppierte Mikroverkalkungen der Befundstufe 4a prägen die invasive Mikrokalkabklärung [15].

Eine Differenzierung maligner Mikrokalkläsionen bezüglich DCIS-Fällen mit Unterscheidung nach dem Kernmalignitätsgrad und invasiven Mammakarzinomen auf Basis des KI-Scores gelingt in der verwendeten Version nicht. Hinsichtlich der Detektionsbedeutung wäre es wünschenswert, dass KI insbesondere das DCIS vom intermediären und hohen Kernmalignitätsgrad sowie das invasive Mammakarzinom verlässlich anzeigt [18]. Unter den KI-falsch negativ bewerteten Läsionen befanden sich vorrangig DCIS-Läsionen (92,9 %, 13/14), allerdings jeden Kernmalignitätsgrades, und ein invasives Mammakarzinom (7,1 %, 1/14). Prospektive Studien zum KI-Einsatz in der mammografischen Befundung werden als notwendig erachtet, unter anderem um die Performance in der Interaktion der Detektion biologisch relevanter, die Brustkrebssterblichkeit beeinflussender Diagnosen im Kontext einer optimalen Rate benigner Abklärungen zu erheben [9, 19].

Die besondere Stärke der vorliegenden Arbeit ist die Läsionsbezogene KI-Bewertung mit hoher Fallzahl. Die Abklärungsdiagnostik sowie die histologische Befundung unterlagen einem hohen Maß an Standardisierung mit Follow-up. Die Studiendaten waren kein Teil des Datensatzes, auf dem das KI-System trainiert wurde.

Als Limitation ist zu benennen, dass das Studiendesign nicht darauf ausgelegt war, eine KI-bezogene Detektion zusätzlicher, Mikrokalk-assoziierter, maligner Läsionen neben den Rückrufläsionen zu prüfen, bzw. die Sensitivität für Mikrokalk-assoziierte Malignome zu steigern, da kein Abgleich mit Intervallkarzinomen für den Zeitraum vorlag. Der Einsatz von KI zur Reduktion benigner Abklärungen auf der Ebene der Befundungsstufen bedarf weiterer Studien inklusive Intervallkarzinomen [20]. Eine retrospektive Studie zeigte, dass bis zu 50,9 % der Intervallkarzinome bereits zum Zeitpunkt des Screenings mit KI-Information detektiert werden können [21]. Zudem wurde weder das Ausmaß zusätzlicher KI-falsch positiver Mikrokalkläsionen noch dessen prospektiver Einsatz mit automatisierter Läsionsanzeige geprüft. Eine Übertragbarkeit der Ergebnisse auf ein anderes diagnostisches Setting könnte gegeben sein, wenn die diagnostischen Eingangsvoraussetzungen, wie ein Ausschluss assoziierter Herde, vergleichbar wäre.

Zusammengefasst konnte mittels KI in dem gewählten Setting ein zur befunderabhängigen Bewertung vergleichbarer PPV3 für Mikrokalkläsionen gesamt und pro Biopsie-indizierende Kategorie erzielt werden. Eine sich auf den PPV3 auswirkende Minderung Screening-negativer Mikrokalkabklärungen ohne Brustkrebsnachweis zeigte die KI-Anwendung nicht. Die differenzierte Betrachtung der KI-Performance pro Befundstufe ergab bei Einsparung falsch positiver Biopsien insbesondere falsch negative Bewertungen in der Gruppe mit geringstem Malignitätsverdacht 4a. Eine Score-spezifische histologische Läsionsdifferenzierung lieferte das System in der vorliegenden Studie nicht.

KLINISCHE RELEVANZ

Die angewandte KI erreicht über alle radiologischen Befundstufen im Vergleich zur menschlichen Bewertung keine Steigerung der positiven prädiktiven Werte für die invasive Mikrokalkabklärung.

Insbesondere bei geringstem radiologischen Verdachtsgrad erscheint eine dezidierte menschliche Bewertung sinnvoll aufgrund eines potenziell höheren Risikos einer KI-falsch negativen Bewertung als in den suspekteren Befundstufen.

Funding

EU INTERREG V A Programm Deutschland-Niederlande; Projekt InMediValue 122 207

Interessenkonflikt

Die Autorinnen/Autoren geben an, dass kein Interessenkonflikt besteht.

Literatur

[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 microcalcification 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 microcalcifications 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