

Statusartikel

Ugeskr Læger 2020;182:V06200423

Kunstig intelligens til cancerdiagnostik i brystkræftscreening

Mohammad Talal Elhakim^{1, 2}, Ole Graumann^{1, 2}, Lisbet Brønros Larsen^{1, 2}, Mads Nielsen³ & Benjamin Schnack Rasmussen^{1, 2}

1) Radiologisk Afdeling, Odense Universitetshospital, 2) Radiologisk Forsknings- og Innovationsenhed, Syddansk Universitet, 3) Datalogisk Institut, Københavns Universitet

Ugeskr Læger 2020;182:V06200423

HOVEDBUDSKABER

- Kunstig intelligens (AI) kan potentielt effektivisere brystkræftdiagnostik i screening.
- Deep learning-metoder har vist performance på menneskeligt niveau, og kommercielle AI-systemer er allerede i brug på verdensplan.
- Implementering af AI i dansk screeningspraksis forudsætter klinisk validering.

Brystkræft er den hyppigst forekommende cancer hos kvinder i Danmark med den højeste dødelighed efter lungekræft [1]. Organiseret, populationsbaseret mammografiscreening er implementeret med det formål at reducere mortaliteten af brystkræft gennem tidlig detektion, og efter udrulningen af det nationale screeningsprogram i 2007-2010 har antallet af deltagere været støt stigende [2]. Pga. en øget byrde og sparsomme radiologressourcer er området præget af et stort arbejdspress. Desuden er mammografiscreening i litteraturgennemgange af internationale studier forbundet med et vist antal falsk positive [3] samt risiko for overdiagnosticering og dermed potentiel overbehandling [4], om end danske tal efter flere screeningsrunder indikerer et forholdsvist lavere niveau [5].

Kunstig intelligens (AI) har i de senere år tiltrukket sig tiltagende interesse som et potentielt værktøj, der kan øge effektiviteten og kvaliteten inden for billeddiagnostik. Et nyligt publiceret studie fra Google Health i tidsskriftet Nature har bl.a. medført stor medieomtale for konklusionen om, at AI overgår radiologer i at diagnosticere brystkræft [6]. Formålet med denne artikel er at: 1) præsentere læseren for den nyeste viden om AI til cancerdiagnostik i brystkræftscreening, 2) gennemgå metodologiske problemstillinger i den eksisterende litteratur om AI til mammografi og 3) diskutere mulighederne samt udfordringerne i at implementere AI-løsninger i klinisk praksis.



Screeningsbilleder dobbeltgranskes uafhængigt af to specialiserede brystradiologer.

NUVÆRENDE PRAKSIS I BRYSTKRÆFTSCREENING

Alle danske kvinder i aldersgruppen 50-69 år inviteres hvert andet år til screening for brystkræft. Undersøgelsen består af en full-field digital mammografi, som er en standardiseret højkvalitetsrøntgenoptagelse af hvert bryst i to projektioner. Screeningsbillederne læses ved dobbelt uafhængig granskning af rutinerede mammaradiologer, og i tilfælde af uenighed foretages en tredje granskning mhp. konsensus. Screeningen afsluttes enten ved, at kvinden fortsætter i screeningsprogrammet, eller at hun bliver genindkaldt til nærmere udredning med klinisk mammografi (palpation, supplerende billeddiagnostik og evt. nålebiopsi). Digital brysttomosyntese (DBT), som er en tredimensionel mammografi, foretages i udvalgte tilfælde. Videre udredning og behandling foregår i multidisciplinært organiseret pakkeforløb.

Kvaliteten af screeningsindsatsen monitoreres løbende af Dansk Kvalitetsdatabase for Mammografiscreening [2] ud fra en række organisatoriske krav og kliniske kvalitetsindikatorer iht. europæiske retningslinjer [7] og nationale guidelines [8]. Opgørelserne fra de seneste screeningsrunder viser, at den nationale mammografiscreening har en høj lægefaglig standard, men fortsat ressourcemæssige udfordringer [2].

COMPUTERBASEREDE LØSNINGER TIL MAMMOGRAFI

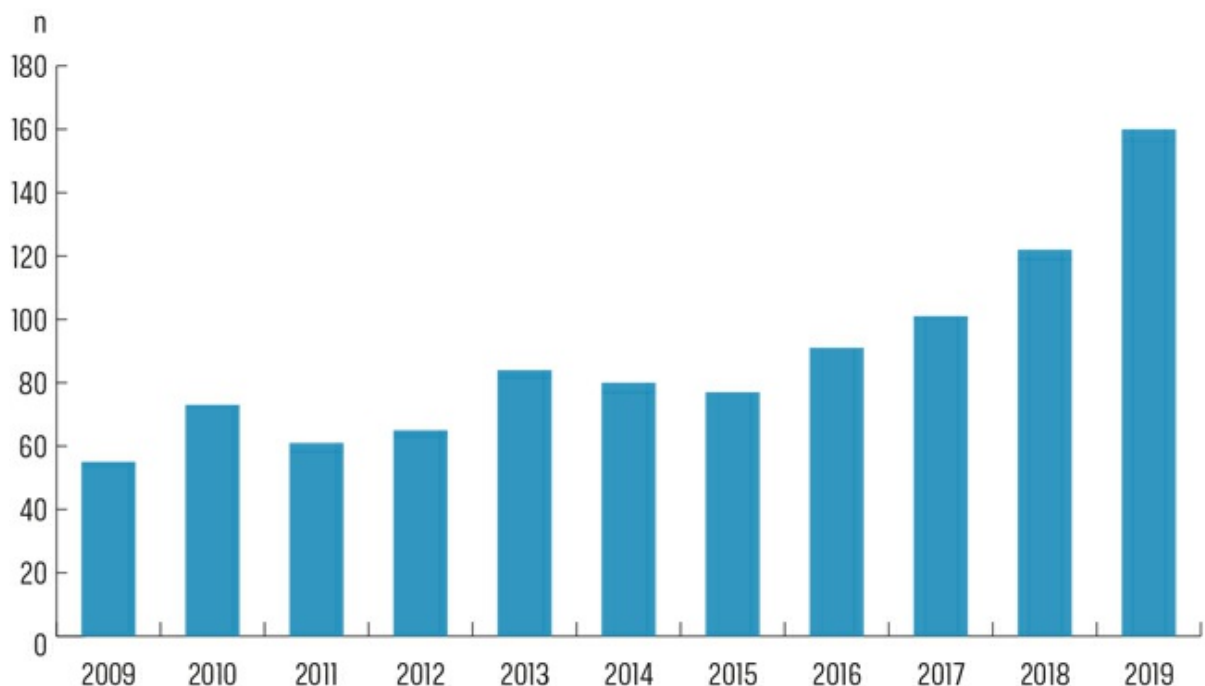
De første computerassisterede detektions (CAD)-systemer blev i slutningen af 1990'erne lanceret som et støtteværktøj til at bistå radiologer i brystkræftdiagnostik med store forventninger om at reducere oversete cancere samt falsk positive. Med transitionen fra konventionel røntgen på fotografisk film til digital mammografi har CAD inden for de seneste 20 år opnået bred popularitet særligt i USA, hvor der bruges enkeltgranskning i screening [9]. En overvægt af studier peger dog på, at CAD forringer den diagnostiske kvalitet, med bl.a. afledende falsk positive markeringer, uden reel effektivisering af arbejdsgangen [10]. I Europa har CAD derfor ikke vundet stort indpas, og dobbeltgranskning anbefales stadig som standard [7].

De traditionelle CAD-systemer brugte »primitive« AI-algoritmer, som krævede manuel programmering og annotering af de enkelte vævsstrukturer. Moderne algoritmer til billedanalyse gør imidlertid brug af machine learning (ML)-teknikker, overvejende af typen deep learning (DL) med store neurale netværk, som muliggør automatisk oplæring alene ud fra store mængder data. Potenseret af nye grafikprocessorerheder med stor ydeevne er disse algoritmer de klassiske langt overlegne, både hvad angår dataprocessering og mønstergenkendelse [11].

EN KRITISK EVALUERING AF DEEP LEARNING-SYSTEMER TIL MAMMOGRAFI

Med frembruddet af den nyeste DL-teknologi fra omkring 2012 er antallet af publicerede videnskabelige artikler om AI-metoder til mammografi mere end fordoblet (Figur 1). Langt hovedparten er udført i et forskningsmiljø, hvori algoritmerne er trænet og validerede på begrænsede datasæt. Effektmålene for cancerdetektion består overvejende af sensitivitet og specificitet afbildet i en receiver operating characteristic-kurve. Trods flotte resultater på op til 98% i arealet under kurven (AUC) [12] giver den tekniske performance i sig selv ikke grundlag for sammenligning med radiologers performance. Desuden er datasættene ofte urealistisk homogene og har en ikkerepræsentativ prævalens af brystkræft.

FIGUR 1 / Søgningresultatet for antallet af publicerede videnskabelige artikler om kunstig intelligens og mammografi i perioden 2009-2019 på Web of Science med søgestrengen: TS=(((mammography OR breast cancer screening) AND (deep learning OR machine learning OR artificial intelligence OR convolutional neural network OR deep neural network OR computer assist OR CAD OR Computer-aided detection OR computer-aided diagnosis)) NOT (MRI OR Ultrasound OR Thermography)). Søgningen blev foretaget d. 2. juni 2020.



Kunstig intelligens versus radiologer

I de seneste to år er der lavet en række komparative undersøgelser [6, 13-21], hvor AI-systemers performance

sammenlignes med radiologers, og flere kommercielt tilgængelige AI-systemer til mammografi er godkendt til klinisk brug (Tabel 1) med stigende anvendelse på verdensplan. Undersøgelserne er foretaget enten på en retrospektiv kohorte med radiologens vurdering og/eller histopatologi som reference eller som et reader study med AI-systemet sat op imod radiologer over for et udvalgt datasæt med en høj cancerprævalens. I alle studier rapporteres der om en forbedret radiologperformance med AI-systemet som støtteværktøj og/eller en noninferiør performance af AI-systemet som selvstændig gransker udtrykt ved AUC – i nogle tilfælde endda bedre end radiologerne. I enkelte studier har man herudover via simulationsanalyser fundet en væsentlig optimering af arbejdsgangen, idet man ved at erstatte den ene gransker med AI-systemet kan reducere arbejdsbyrden med op til 50%.

TABEL 1 / Deep learning-baserede kunstig intelligens (AI)-systemer til cancerdetektion på mammografi, som er enten kommercielt tilgængelige eller undersøgt i publiceret peer review-materiale efter 2018, hvori performance er sammenholdt med radiologer som reference. Listen er ikke udtømmende for eksisterende systemer.

Udvikler	AI-system	Kommercielt tilgængeligt	Peer review	Løsning til DBT	Løsning til andet end mammografi	CE-mærket/ FDA-godkendt
Zebra Medical Vision, Inc.	AI1	Ja	Nej ^a	Nej	Ja ^b	Ja/ja ^a
CureMetrix, Inc.	cmTriage	Ja	Nej	Nej	Nej	Nej/ja
	cmAssist	Nej	Ja [13]	Nej	Nej	Nej/nej
Hologic, Inc.	Genius AI	Ja	Nej	Ja	Nej	Nej/nej
Lunit, Inc.	Lunit INSIGHT MMG	Ja	Ja [14]	Nej	Nej	Ja/nej ^c
Kheiron Medical Technologies	Mia	Ja	Nej	Nej	Nej	Ja/nej ^c
iCAD, Inc.	ProFound AI	Ja	Ja ^d [15]	Ja	Nej	Ja/ja ^d
ScreenPoint Medical BV	Transpara	Ja	Ja [16-18]	Ja	Nej	Ja/ja
<i>Forskningsystemer uden et produkt navn</i>						
Google Health	-	Nej	Ja [6]	Nej	Nej	Nej/nej
IBM Research	-	Nej	Ja [19]	Nej	Nej	Nej/nej
MIT	-	Nej	Ja [20]	Nej	Nej	Nej/nej
New York University	-	Nej	Ja [21]	Nej	Nej	Nej/nej

AI1 = All-In-One; CE = Conformité Européenne; DBT = digital brysttomosyntese; FDA = U.S. Food and Drug Administration; IBM = International Business Machines Corporation; Mia = Mammogram Intelligent Assessment; MIT = Massachusetts Institute of Technology.

a) Kun til de øvrige løsninger i AI1.

b) AI1 kan også vurdere osteoporotiske sammenfald på CT samt akutte forandringer på røntgenundersøgelse af thorax og CT af cerebrum.

c) Ansøgt.

d) Kun til DBT.

Kliniske valideringsstudier som en videnskabelig guldstandard

Man er med komparative forsøg kommet et væsentligt skridt nærmere evidensbaseret implementering af AI-løsninger til mammografi, men proof-of-concept studier kan og må ikke stå alene. Resultaterne har begrænset generaliserbarhed til klinisk screeningspraksis og til andre populationer, hvilket et japansk reader study tyder på med dårligere performance af et kommercielt europæisk udviklet AI-system end radiologer [22]. Dette understreger vigtigheden af såvel retrospektive som prospektive studier og randomiserede kontrollerede forsøg, hvor AI-systemer valideres i lokal klinisk kontekst – en generel mangelvare inden for billeddiagnostik [23], hvor forskningen i AI ikke lever op til en høj standard [24].

Flere metodologiske standarder til rapportering af medicinsk forskning forsøges adapteret AI- og ML-metoder. Ansete internationale guidelines som TRIPOD, CONSORT, SPIRIT og STARD er undervejs med udvidelserne TRIPOD-ML, CONSORT-AI, SPIRIT-AI og STARD-AI [25]. Det amerikanske tidsskrift Radiology har for nylig udgivet en række anbefalinger, som i mellemtiden skal hjælpe med evaluering af forskningen i AI [26]. Disse er opsummeret i Tabel 2 med reference til de studier [6, 13-21], der opfylder hver enkel anbefaling. Her bemærkes, at det tidligere omtalte Google studie af McKinney et al [6], som de fleste andre, hverken har testet deres model

på et eksternt datasæt eller gjort algoritmen offentligt tilgængelig, hvilket svækker reproducerbarheden.

TABEL 2 / Anbefalinger vedrørende forskning i kunstig intelligens (AI) og machine learning inden for billeddiagnostik til forfattere, reviewere og læsere. Adapteret og opsummeret fra [26]. Der er refereret til de studier, der opfylder den enkelte anbefaling.

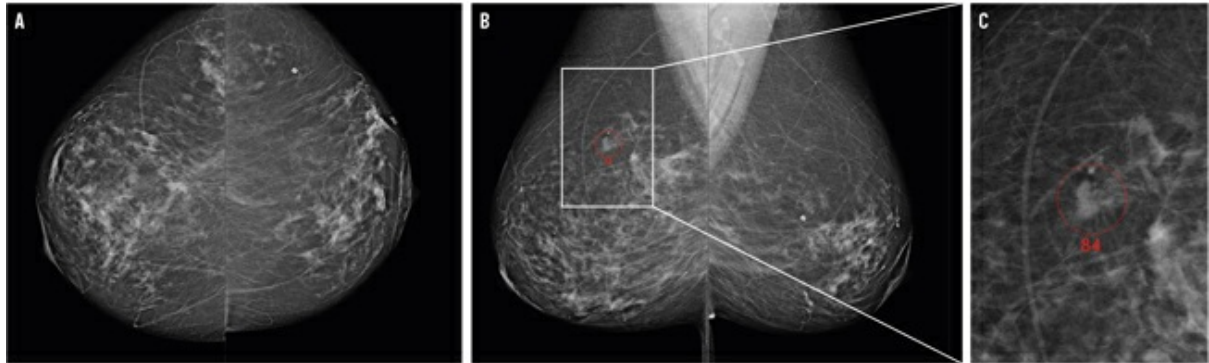
Anbefaling	Referencer
<p><i>Definer alle 3 datasæt (til træning, validering og test)</i> Data til træning, validering og test bør optimalt set opdeles i 3 uafhængige sæt for at udvikle en robust og brugbar AI-model</p>	<p>Akselrod-Ballin et al, 2019 [19] Kim et al, 2020 [14] McKinney et al, 2020 [6] Wu et al, 2020 [21] Yala et al, 2019 [20]</p>
<p><i>Brug et eksternt datasæt til test</i> AI-modeller har en tendens til at virke bedre på kendte end ukendte datasæt For at sikre generaliserbarhed bør den endelige test med tilhørende statistisk analyse foretages på et eksternt datasæt, f.eks. fra en anden institution</p>	<p>Conant et al, 2019 [15] Kim et al, 2020 [14] Rodriguez-Ruiz et al, 2019 [16-18]</p>
<p><i>Brug billeder fra flere leverandører i alle faser</i> Billeder fra forskellige leverandører indeholder teknologiske forskelle som registreres af AI-algoritmerne Derfor er leverandør-specifikke algoritmer mindre interessante end leverandør-uafhængige set i et globalt perspektiv</p>	<p>Kim et al, 2020 [14] Rodriguez-Ruiz et al, 2019 [16-18] Wu et al, 2020 [21]</p>
<p><i>Begrund størrelsen af de 3 datasæt</i> Antallet af billeder der ligger til grund for udviklingen af AI-algoritmen afhænger af applikationen og bør sammenholdes med performance Den ideelle størrelse af et datasæt kendes ofte ikke men som tommelfingerregel kræves mange 1.000</p>	
<p><i>Træn algoritmen op med en anerkendt referencestandard</i> Datasættet til træning bør være annoteret og grundigt vurderet af flere eksperter på baggrund af den mest anerkendte standard inden for feltet, f.eks. kan histopatologi godt benyttes som reference</p>	<p>Akselrod-Ballin et al, 2019 [19] Kim et al, 2020 [14] McKinney et al, 2020 [6] Wu et al, 2020 [21] Yala et al, 2019 [20]</p>
<p><i>Beskriv evt. dataforberedelse til algoritmen</i> Enhver selektion eller tilpasning af data til AI-algoritmen kan påvirke radiologens forståelse af AI-modellens tolkning, f.eks. hvis den kun præsenteres for udvalgte snit fra en CT frem for alle billederne</p>	<p>Wu et al, 2020 [21] Yala et al, 2019 [20]</p>
<p><i>Sammenhold performance med førende radiologer</i> Det er bedre at evaluere AI-modellen op imod eksperterne på området frem for generalisterne, alternativt histopatologisk svar</p>	<p>Akselrod-Ballin et al, 2019 [19] Conant et al, 2019 [15] Kim et al, 2020 [14] McKinney et al, 2020 [6] Rodriguez-Ruiz et al, 2019 [16-18] Watanabe et al, 2019 [13] Wu et al, 2020 [21] Yala et al, 2019 [20]</p>
<p><i>Demonstrer, hvordan algoritmen træffer beslutninger</i> Arealet under kurven for performance har begrænset klinisk anvendelighed. Derfor bør beslutningsprocessen beskrives med f.eks. heat maps, farvekoder eller sandsynlighedsmaps Hvis det er muligt at justere skæringsværdier for sensitivitet og specificitet for forskellige features skal dette også være synligt</p>	<p>Akselrod-Ballin et al, 2019 [19] Kim et al, 2020 [14] McKinney et al, 2020 [6] Wu et al, 2020 [21]</p>
<p><i>Algoritmen bør være offentligt tilgængelig</i> For at muliggøre uafhængig reproducerbarhed bør AI-algoritmer, også kommercielle, være offentligt tilgængelige på f.eks. onlinesoftwareplatforme så de kan komme så mange patienter som muligt til gode</p>	<p>Akselrod-Ballin et al, 2019 [19] Wu et al, 2020 [21] Yala et al, 2019 [20]</p>

KLINISK IMPLEMENTERING – MULIGHEDER OG UDFORDRINGER

Størstedelen af deltagerne i en screeningspopulation er raske. Hvis AI-systemer har en tilstrækkelig høj sensitivitet og negativ prædiktiv værdi, er triagering en oplagt anvendelsesmåde i mammografiscreening [18, 20] (Figur 2), hvor det i en europæisk kontekst med dobbeltgranskning vil kunne erstatte den ene gransker. Hvis man bruger et triageringsværktøj som første gransker, kan man differentiere undersøgelser med og uden suspekter forandringer og sende dem til hhv. dobbeltgranskning og enkeltgranskning, så kun mammografierne i højrisikogruppe vurderes af to radiologer. Til DBT, der er mere tidskrævende at gennemse, men på sigt kan få en større rolle i mammografiscreening, kan AI bruges som et støtteværktøj til reduktion af tidsforbruget og

forbedring af performance [15].

FIGUR 2 / Kunstig intelligens (AI)-systemer til cancerdetektion på mammografi kan anvendes som et interaktivt beslutningsstøtteværktøj, en automatisk selvstændig gransker og/eller et triageringsværktøj til at sortere billederne efter mistanke om cancer. Her ses en digital screeningsmammografi på en 53-årig kvinde af begge bryster optaget i to projektioner: kraniokaudal (A) og mediolateral oblique (B). Som støtteværktøj kan et AI-system markere en given læsion/masse med f.eks. en rød ring med tilhørende malignitetsscore. Her vurderede systemet et biopsiverificeret invasivt duktalt mamma-karcinom med en score på 84 ud af 100 (100 angiver størst sandsynlighed for cancer). C. Forstørrelse af læsionen/massen.



Figuren er gengivet med tilladelse fra ScreenPoint Medical BV.

En af de potentielle faldgruber ved kun at vægte den diagnostiske præcision er, at man finder flere ikkeklinisk betydende cancere. Det ultimative mål bør være kun at identificere klinisk betydende cancer og således undgå overdiagnosticering. Dette forudsætter, at man i AI-forskningen ser ud over præcision og undersøger prædiktive og prognostiske markører. DL-metoder har f.eks. vist en forbedret risikoestimering for udvikling af brystkræft ift. eksisterende modeller [27]. Risikostratifikation kan således være med til at understøtte et individualiseret screeningsprogram ud fra den enkelte kvindes risikofaktorer, såvel klinisk som billeddiagnostisk.

Ovennævnte muligheder bør valideres i langsigtede prospektive studier, som samtidig kan sikre en praktisk ladsiggørlig integration af AI-løsninger i den eksisterende radiologiske arbejdsgang. De kommercielt tilgængelige systemer muliggør direkte integration med den radiologiske arbejdsstation eller en cloud-baseret løsning til reeltidsanalyse. Dog er det svært at gennemskue konsekvensen ved køb af en AI-løsning til et givent sundhedstilbud, da der os bekendt endnu ikke eksisterer en fuldendt evalueringsmodel til implementering af AI i klinikken.

JURA OG ETIK

Datasikkerhed er et centralt emne ift. datadrevet teknologi som AI, for hvilket persondataforordningen ligeledes er gældende. Eftersom lovgivningen endnu ikke tager højde for AI inden for sundhedsteknologi, er korrekt evaluering og især gentræning af AI-systemerne på nogle områder begrænset. På et politisk plan er der i Danmark udvist vilje til at følge med udviklingen, hvilket kom til udtryk i regeringens nationale strategi for kunstig intelligens i foråret 2019 [28]. Tilsvarende tendenser ses i europæisk sammenhæng. Det er først for nylig, at krav om monitorering af medicinske AI-systemer efter implementering er italesat af den amerikanske Food and Drug Administration (FDA) [29]. FDA-godkendelse og Conformité Européenne (CE)-mærkning er i hhv. USA og EU en lovgivningsmæssig markedsgodkendelse, som garanterer en vis sikkerhed og kvalitet af kommercielt medicinsk udstyr. Nogle AI-systemer er godkendte til mammografi uden peer review eller alene på baggrund af data til øvrig diagnostik, bl.a. røntgen af thorax (Tabel 2), hvorfor mærkninger i det hele taget ikke forudsætter betydelig evidens og således har begrænset videnskabelig betydning. Det er vigtigt at understrege, at opbygningen af et AI-system kan variere, da et system f.eks. kan være statisk (ikke-lærende) eller adaptivt (lærer

undervejs), og til vores viden er der ikke fundet nogen løsning på, hvordan et adaptivt system skal godkendes.

Implementering af AI rejser i forlængelse heraf vigtige etiske spørgsmål om bl.a. gennemsigtighed og ansvar ved fejl ift. algoritmernes autonomi. I april 2019 udgav en uafhængig ekspertgruppe under Europa-Kommissionen etiske retningslinjer for udvikling og anvendelse af pålidelig AI [30]. Lovgivning og internationale retningslinjer er nødvendige for at sikre etiske rammer omkring implementering af AI.

PERSPEKTIVER

Radiologien har historisk været god til at tilpasse sig teknologisk udvikling, og vi ser frem mod en ny tid bl.a. med brystkræftscreening drevet af AI. Dansk Radiologisk Selskab har i efteråret 2019 nedsat en national gruppe netop med fokus på billeddiagnostisk AI for bl.a. at sætte dagsordenen i dansk klinisk sammenhæng.

AI udgør et lovende element i at aflaste screeningsprogrammerne uden at gå på kompromis med kvaliteten. Den danske mammografiscreening har med sine organiserede databaser god grobund for klinisk validering af AI-løsninger forud for implementering. Generelt skal processen med arbejdet omkring AI være evidensbaseret og præget af en holistisk tilgang, der omfavner de mange aspekter af AI. Altafgørende er, at udokumenterede gevinster og innovation ikke går forud for ordenligt evaluerede sundhedsteknologier, hvor de personale- og ikke mindst patientmæssige konsekvenser kan være store. For at komme i mål er det essentielt, at vi samarbejder på tværs af fagligheder (læger, dataloger, sundheds- og regulatoriske myndigheder mv.) og sørger for, at solide data er tilgængelig til at bistå evaluering og ikke mindst den fremtidige udvikling af AI med Danmark som foregangsland.

KORRESPONDANCE: *Mohammad Talal Elhakim*. E-mail: mte@rsyd.dk

ANTAGET: 17. juli 2020

PUBLICERET PÅ UGESKRIFTET.DK: 17. august 2020

INTERESSEKONFLIKTER: Forfatterens ICMJE-formularer er tilgængelige sammen med artiklen på Ugeskriftet.dk

LITTERATUR: Findes i artiklen publiceret på Ugeskriftet.dk

SUMMARY

Artificial intelligence for cancer detection in breast cancer screening

Mohammad Talal Elhakim, Ole Graumann, Lisbet Brønros Larsen, Mads Nielsen & Benjamin Schnack Rasmussen

Ugeskr Læger 2020;182:V06200423

Artificial intelligence (AI) has the potential to increase quality and efficiency of breast cancer screening. Recent studies have provided comparative data on AI versus human performance in cancer detection with encouraging results, and commercially available AI systems are used worldwide as a clinical tool for mammography screening. There are, however, a number of methodological concerns in relation to the evaluation of AI systems. This review discusses these aspects as well as the opportunities and challenges of clinical validation and implementation in breast cancer screening practice.

LITTERATUR

1. Association of the Nordic Cancer Registries. NORDCAN databasen. <https://nordcan.iarc.fr/en> (2. jun 2020).
2. Regionernes Kliniske Kvalitetsudviklingsprogram. Dansk Kvalitetsdatabase for Mammografiscreening. www.rkkp.dk (2. jun 2020).

3. Hofvind S, Ponti A, Patnick J et al. False-positive results in mammographic screening for breast cancer in Europe: a literature review and survey of service screening programmes. *J Med Screen* 2012;19(suppl 1):S57-S66.
4. Marmot MG, Altman DG, Cameron DA et al. The benefits and harms of breast cancer screening: an independent review. *Br J Cancer* 2013;108:2205-40.
5. Lynge E, Bak M, von Euler-Chelpin M et al. Outcome of breast cancer screening in Denmark. *BMC Cancer* 2017;17:897.
6. McKinney SM, Sieniek M, Godbole V et al. International evaluation of an AI system for breast cancer screening. *Nature* 2020;577:89-94.
7. Perry N, Broeders M, de Wolf C et al. European guidelines for quality assurance in breast cancer screening and diagnosis. 4th ed. Luxembourg, 2006.
8. Danish Breast Cancer Group. Retningslinjer for billeddiagnostisk udredning, 2013. www.dbcg.dk (2. jun 2020).
9. Keen JD, Keen JM, Keen JE. Utilization of computer-aided detection for digital screening mammography in the United States, 2008 to 2016. *J Am Coll Radiol* 2018;15:44-8.
10. Lehman CD, Wellman RD, Buist DSM et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med* 2015;175:1828-37.
11. Ranschaert ER, Morozov S, Algra PR. Artificial intelligence in medical imaging. 1st ed. Springer, 2019.
12. Houssami N, Kirkpatrick-Jones G, Noguchi N et al. Artificial intelligence (AI) for the early detection of breast cancer: a scoping review to assess AI's potential in breast screening practice. *Expert Rev Med Devices* 2019;16:351-62.
13. Watanabe AT, Lim V, Vu HX et al. Improved cancer detection using artificial intelligence: a retrospective evaluation of missed cancers on mammography. *J Digit Imaging* 2019;32:625-37.
14. Kim H-E, Kim HH, Han B-K et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit Health* 2020;2:e138-e148.
15. Conant EF, Toledano AY, Periaswamy S et al. Improving accuracy and efficiency with concurrent use of artificial intelligence for digital breast tomosynthesis. *Radiol Artif Intell* 2019;1:e180096.
16. Rodriguez-Ruiz A, Krupinski E, Mordang JJ et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. *Radiology* 2019;290:305-14.
17. Rodriguez-Ruiz A, Lang 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;111:916-22.
18. Rodriguez-Ruiz A, Lang K, Gubern-Merida A et al. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? *Eur Radiol* 2019;29:4825-32.
19. Akselrod-Ballin A, Chorev M, Shoshan Y et al. Predicting breast cancer by applying deep learning to linked health records and mammograms. *Radiology* 2019;292:331-42.
20. Yala A, Schuster T, Miles R et al. A deep learning model to triage screening mammograms: a simulation study. *Radiology* 2019;293:38-46.
21. Wu N, Phang J, Park Y et al. deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Trans Med Imaging* 2020;39:1184-94.
22. Sasaki M, Tozaki M, Rodriguez-Ruiz A et al. Artificial intelligence for breast cancer detection in mammography: experience of use of the ScreenPoint Medical Transpara system in 310 Japanese women. *Breast Cancer* 12. feb 2020.
23. Nagendran M, Chen Y, Lovejoy CA et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020;368:m689.
24. Liu X, Faes L, Kale AU et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 2019;1:e271-e297.
25. Sounderajah V, Ashrafian H, Aggarwal R et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. *Nat Med* 2020;26:807-8.
26. Bluemke DA, Moy L, Bredella MA et al. Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers—from the radiology editorial board. *Radiology* 2020;294:487-9.
27. Dembrower K, Liu Y, Azizpour H et al. Comparison of a deep learning risk score and standard mammographic density score for breast cancer risk prediction. *Radiology* 2020;294:265-72.

28. Regeringen. National strategi for kunstig intelligens, 2019. <https://www.regeringen.dk> (2. jun 2020).
29. U.S. Food and Drug Administration. Artificial intelligence and machine learning in software as a medical device, 2020. <https://www.fda.gov> (2. jun 2020).
30. European Commission. Ethics guidelines for trustworthy AI, 2019. <https://ec.europa.eu> (2. jun 2020).