Så kan AI valideras för klinisk implementering

Fredrik Strand, docent i radiologi, röntgenläkare, Karolinska institutet; Karolinska universitetssjukhuset
● fredrik.strand@ki.se

Sophia Zackrisson, professor i radiologi, överläkare, Lunds universitet; Skånes universitetssjukhus

Håkan Gustafsson, docent i medicinsk strålningsfysik; biträdande föreståndare, Centrum för medicinsk bildvetenskap och visualisering (CMIV); AI-koordinator, Region Östergötland

Under senare år har tillgången till medicintekniska produkter baserade på artificiell intelligens (AI) ökat snabbt inom hälso- och sjukvården, och bara inom radiologi finns i nuläget över 200 CE-märkta produkter [1]. AI har börjat nå en tillräcklig nivå för självständiga bedömningar av röntgenbilder, och resultat från retrospektiva och prospektiva studier inom mammografi visar på tydlig klinisk nytta [2-5].

Medicinska tillämpningar av AI är ett område som är under mycket snabb utveckling, och i nuläget ställer certifierande myndigheter relativt låga krav för att en produkt baserad på AI ska kunna CE-märkas som en medicinteknisk produkt [6]. Det är välkänt att även etablerade AI-produkter kan ge felaktiga svar om de används på en annan population än de tränats på [7, 8], och kraven för att en AI-baserad medicinteknisk produkt ska kunna CE-märkas (och/eller FDA-godkännas) baseras i första hand på att man kan visa att den är lika bra som en redan existerande produkt eller att den tillför klinisk nytta [6, 9]. Det förekommer att AI-produkter enbart har testats retrospektivt och med data från ett eller ett fåtal sjukhus.

Vi som står patienterna närmast måste i nuläget ta ansvar för vilka AI-system vi väljer att använda och hur vi använder dem. Att utan eftertanke hoppa på pilotstudier som ofta erbjuds av AI-företagen kan utgöra en risk för patientsäkerheten, eftersom studiepopulationen ofta blir alltför liten för statistiskt hållbara slutsatser. En gedigen validering av AI kan vara övermäktig för den enskilde läkaren och kliniken. Förmågan att systematiskt utvärdera, kvalitetssäkert implementera och löpande övervaka AI-systemen behöver finnas nära kopplad till sjukvården. När det gäller det första steget, utvärdering, kan det sägas bestå av två delar. För att få en uppfattning om den diagnostiska träffsäkerheten hos en AI-algoritm kan man låta den analysera existerande mammografibilder med kända cancerutfall – en så kallad retrospektiv validering. För att sedan förstå hur den diagnostiska potentialen tas till vara i praktiken och hur diagnostiker samspelar med AI-systemen behövs prospektiva interventionsstudier, där AI får påverka de kliniska besluten. Det kan hända att förutsättningarna i en publicerad prospektiv studie inte är desamma som på ett sjukhus där du arbetar, och därför kan det även behövas validering utifrån lokala förutsättningar. Inom mammografiscreening väntas flera prospektiva studier snart rapportera resultat. Ett abstrakt avseende en interventionsstudie på Capio S:t Görans sjukhus presenterades på RSNA:s (Radiological Society of North America) kongress 2022 och ett annat avseende en interventionsstudie i Region Skåne presenterades på kongressen European Congress of Radiology 2023. De presenterade resultaten bekräftade de retrospektiva resultaten att dagens AI-system tillräckligt säkert kan bedöma förekomst av cancermisstänkta förändringar inom mammografiscreening.

Projektet VAI-B, Validering av AI inom bröstradiologi, startades under hösten 2021 som ett svar på behovet av validering av AI-algoritmer inom bröstradiologi och framför allt mammografiscreening [10]. Vi erhöll finansiering av Vinnova och Regionala cancercentrum i samverkan. Vetenskaplig ledare för projektet är Fredrik Strand, docent på Karolinska institutet, och Sophia Zackrisson, professor vid Lunds universitet. Den aktiva projektgruppen består av medverkande forskare, projektledare, IT-arkitekt, forskningsingenjör, biostatistiker och partneransvarig. Tanken är att erbjuda systematiska utvärderingar av AI-system genom att låta dem bearbeta existerande mammografibilder och jämföra med facit i form av bröstcancerdiagnos inom en bestämd uppföljningstid efter undersökningen. Hittills har 3 regioner och 3 AI-företag valt att delta: Östergötland, Södermanland och Västmanland samt Vara (Tyskland), Therapixel (Frankrike) och Lunit (Sydkorea). Från de deltagande regionerna samlar vi in uppgifter om vilka kvinnor som deltagit i screening mellan 2008 och 2021. Därefter länkas dessa med det nationella kvalitetsregistret för bröstcancer för att avgöra vilka som erhållit diagnos. Regionerna överför pseudonymiserade röntgenbilder för alla diagnostiserade individer samt ett slumpmässigt urval av friska.

Nyligen publicerades en artikel där vi beskriver hur plattformen rent tekniskt är uppbyggd [11]. Den är utformad som en hybrid med en molnlösning för röntgenbilder och AI-system samt en lokal server på Karolinska institutet för utfallsdata och analyser. De siffror och bilder som AI-systemen producerar i molnlösningen skickas till den lokala servern för vidare bearbetning. Valideringen utförs genom att biostatistiker beräknar hur väl AI-systemen presterar diagnostiskt samt hur de påverkas av variationer i populationen och olika förutsättningar i bildtagningen (till exempel att bilder kommer från mammografiutrustningar av olika modeller). I pilotstudien undersöker vi hur AI fungerar i den initiala mammografigranskningen, där det normalt sett är två röntgenläkare som granskar varje undersökning och flaggar för att gå vidare till konsensusdiskussion om de ser något misstänkt.

Hur väl AI-systemen presterar kan mätas på olika sätt. Att mäta den diagnostiska förmågan med ett enda mått är oftast inte tillräckligt. Exempelvis finns måttet träffsäkerhet (balanced accuracy), vilket motsvarar andelen av AI-systemets bedömningar som är korrekta. Problemet är att prevalensen av bröstcancer vid screening är låg: endast cirka 0,5 procent av under-

HUVUDBUDSKAP

- Företag marknadsför över 200 mjukvaror för artificiell intelligens (AI) inom radiologi.
- Vårdgivare behöver ha möjlighet till opartisk och kvalitetssäkrad utvärdering.
- Projektet Validering av AI inom bröstradiologi (VAI-B) startades för att möta detta behov, initialt inom mammografiscreening.

TABELL 1. Sammanställning av mått som mäter olika aspekter av prestanda hos diagnostiska test (till exempel bedömning av mammografiscreening)

	Cancer	Frisk
Positiv (flaggad)	SP	FP
Negativ (ej flaggad)	FN	SN

S = sant, F = falskt, P = positiv, N = negativ

Sensitivitet = SP / (SP + FN)
Detektionsfrekvens för cancer = SP / alla
Specificitet = SN / (SN + FP)
Falskt positiv andel = FP / alla
Positivt prediktivt värde = SP / (SP + FP)

sökningarna motsvarar en diagnos i Sverige. Träffsäkerheten blir då väldigt hög, (99,5 procent) om AI-systemet skulle bedöma alla undersökningar som negativa, det vill säga avsaknad av cancermisstanke, men samtidigt skulle systemet inte bidra med någon som helst nytta. Därför bör man uppvisa åtminstone två olika mått som avspeglar hur AI-systemet bidrar till upptäckten av cancer (sant positiva) och hur det bidrar till onödiga flaggningar (falskt positiva), som leder till både oro och ökad arbetsbelastning vid efterföljande konsensusdiskussion. Mått som avspeglar förmågan till upptäckt av cancer är antingen sensitivitet eller cancerdetektionsfrekvens, medan mått som avspeglar förmågan att undvika falskt positiva bedömningar är specificitet eller falskt positiv andel (Tabell 1). Positivt prediktivt värde (PPV) är ett mått som kan sägas avspegla effektiviteten hos screeningsystemet genom att det visar hur många fall av cancer man upptäcker som andel av antalet flaggningar. Till skillnad från övriga mått är sensitivitet och specificitet inte beroende på prevalensen av en sjukdom, men har nackdelen att det är teoretiskt omöjligt att vara helt säker på vilka kvinnor som saknade cancer vid tidpunkten för mammografin eftersom de som anses friska inte biopseras.

När man jämför AI med röntgenläkare genom retrospektiva data är det viktigt att tänka på att det var röntgenläkarna som avgjorde vilka som återkallades och som kunde få bröstcancerdiagnos. De mammografier där röntgenläkarna inte såg något misstänkt, men där AI i efterhand identifierar misstänkta tecken, föranledde inte någon kallelse för vidare diagnostik. Skulle man utgå från att enbart de kvinnor som diagnostiserades i samband med screeningtillfället har cancer blir det en påtaglig nackdel för AI-systemet jämfört med röntgenläkarna. Därför brukar man använda sig av principen att det räknas som att kvinnan hade bröstcancer vid tidpunkten för mammografin om hon antingen diagnostiserades efter upptäckt vid screeningen eller inom en viss uppföljningstid, ofta 2 eller 3 år.

Även om man visar att ett AI-system har god träffsäkerhet är det viktigt att komma ihåg att det inte nödvändigtvis är sant för olika mindre grupper i populationen. Exempel på variationer i populationen som kan ha betydelse är ålder, mammografisk täthet, geografiskt ursprung och socioekonomisk status. Exempel på variationer i bildtagningen som kan ha betydelse är skillnader i mammografiutrustningen samt byten av hård- eller mjukvara. Man kan även undersöka på totalnivå om AI-systemens bedömningar uppvisar någon systematisk förändring över åren, exempelvis att bilder från 2012 får en annan bedömning än bilder från 2019. Denna typ av förändring över tid kan bero på att AI-systemet är tränat på bilder av en viss kvalitet, som kan förändras över tid på grund av förändringar i röntgenutrustningen, efterbearbetningen eller i rutinerna kring bildtagningen. En ytterligare aspekt, som har belysts, inte minst genom de nya stora språkmodellerna som Chat GPT, är att AI-system kan fungera bra inom ramarna för vad det tränats på, men mindre bra när det ställs inför nya situationer med andra förutsättningar. Det är därför viktigt att inte bara utföra validering vid ett tillfälle utan att ha system för kontinuerlig utvärdering efter att AI tagits i klinisk drift. Ett viktigt begrepp i detta sammanhang är »ML Ops«, som betecknar aktiviteter för att hantera de operativa aspekterna av AI-algoritmer, hur de installeras korrekt och övervakas under drift.

VAI-B bedrivs som ett forskningsprojekt med Karolinska institutet som primär huvudman. Metoder och resultat ägs av denna och kommer att ingå i vetenskapliga publikationer samt arkiveras enligt lagkrav för forskningsdata. De deltagande vårdgivarna och AI-företagen tar del av resultaten både skriftligt och muntligt i den mån de inte utgörs av identifierbara data på patientnivå. Etikprövningsmyndigheten har gett tillstånd för projektet, inklusive ett undantag från individuellt informerat samtycke.

Utöver att plattformen erbjuder en analytisk grund kan det även vara möjligt att använda den när en vårdgivare vill genomföra en upphandling av AI-system, och då kräva att presumtiva säljare låter sina system utvärderas och poängsättas enligt objektiva diagnostiska prestanda på VAI-B-plattformen.

VAI-B är det första exemplet på en valideringsplattform för AI-system. Med tanke på att radiologin, och särskilt bröstradiologin, ligger långt fram i utvecklingen av AI är det sannolikt att liknande valideringsplattformar kommer att växa fram inom andra områden allteftersom godkända AI-system blir tillgängliga. Vi har även ambitionen att bidra till ett europeiskt nätverk av valideringsplattformar via EU-projekt och det planerade initiativet European Health Data Space. En olöst fråga är hur den valideringsfunktion vi utvecklar inom forskningen kan övergå till långsiktig operationell verksamhet. Man kan tänka sig en form av kvalitetsregister för att hålla data säkra, som i sin tur tillhandahåller tjänster direkt till AI-företag, eller möjligen ett gränssnitt som gör det möjligt för tredjepartsföretag att utföra valideringstjänsten utan att ha egen tillgång till data.

● Potentiella bindningar eller jävsförhållanden: Inga uppgivna.

Citera som: Läkartidningen. 2023;120:23065

REFERENSER

1. Diagnostic Image Analysis Group AI for radiology. https://grand-challenge.org/aiforradiology/
2. Salim M, Wåhlin E, Dembrower K, et al. External evaluation of 3 commercial artificial intelligence algorithms for independent assessment of screening mammograms. JAMA Oncol. 2020;6(10):1581-8.
3. Dembrower K, Wåhlin E, Liu Y, et al. Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. Lancet Digit Health. 2020;2(9):E468-74.
4. Lång K, Dustler M, Dahlblom V, et al. Identifying normal mammograms in a large screening population using artificial intelligence. Eur Radiol. 2020;31(3):1687-92.
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(1):88-104.
6. van Leeuwen KG, Schalekamp S, Rutten MJCM, et al. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. Eur Radiol. 2021;31(6):3797-804.
7. Habib AR, Lin AL, Grant RW. The epic sepsis model falls short - the importance of external validation. JAMA Intern Med. 2021;181(8):1040-1.
8. Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med. 2021;181(8):1065-70.
9. Wu E, Wu K, Daneshjou R, et al. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med. 2021;27(4):582-4.
10. Regionalt cancercentrum syd. Nationell valideringsplattform för AI inom mammografiscreening (VAI-B). https://cancercentrum.se/syd/vara-uppdrag/forskning/forsknings--och-innovationsprojekt/tre-projekt-for-att-stodja-anvandandet-av-ai/nationell-valideringsplattform-for-ai-inom-mammografiscreening-vai-b/
11. Cossío F, Schurz H, Engström M, et al. VAI-B: a multicenter platform for the external validation of artificial intelligence algorithms in breast imaging. J Med Imaging (Bellingham). 2023;10(6):061404.