Baraa Alsaadi en Cornelia Schaefer

GLEAMER: AI en botbreuken

Artificiële intelligentie speelt op vele vlakken een toenemende rol, zo ook in de medische wereld. In de radiologie wordt AI ingezet voor beeldreconstructie van CT-en MR-beelden, voor optimalisatie van de workflow, maar ook voor beeldinterpretatie.

CORNELIA SCHAEFER-PROKOP
Radioloog in Meander Medisch Centrum
THOMAS DE KORTE
Radioloog in opleiding in Meander Medisch Centrum
BARAA ALSAADI
Stagiaire bij de afdeling Radiologie van Meander Medisch Centrum

Fractuurdetectie of -uitsluiting is een van de meest frequente radiologische onderzoeken. Vaak komen patiënten direct van de Huisartsenpost of van huis. De radiologiebeelden moeten meteen beoordeeld worden, omdat de patiënt in het geval van een fractuur wordt doorgestuurd naar de SEH voor verdere beleidsbepaling.

De beoordeling van de foto gebeurt door een dienstdoende radioloog die echter nog een heel scala aan andere taken heeft. Daardoor kan het voorkomen dat patiënten langer moeten wachten op de beoordeling van niet-spoedeisende foto's, waardoor de doorstroom bij de Huisartsenpost en de Spoedeisende Hulp hapert.

Er is software op de markt die getraind is via deep learning om fracturen te detecteren op röntgenfoto's van volwassenen en kinderen ouder dan 2 jaar. De AI-software geeft duidelijk aan of er geen of een of meerdere fracturen aanwezig zijn. Ook dubieuze fracturen, waar de AI-analyse dus minder zeker is, worden aangegeven.

Deze AI-tool werd in Meander geïntroduceerd in november 2023. Doel van het hier beschreven onderzoek was zowel te onderzoeken hoe goed het systeem voor fractuurdetectie werkt, maar ook wat de meerwaarde van het systeem zou zijn voor beoordelaars met verschil in ervaring en specialisatie.

Methoden

Röntgenfoto's
Een serie van 660 röntgenfoto's werd samengesteld door een ervaren radioloog en een radioloog in opleiding, beiden deden niet mee aan de beoordeling van de foto's. Voor zeven lichaamsgebieden werden foto's geselecteerd: schouder, elleboog, pols/hand, bekken, knie en enkel/voet. De helft van de foto's bevatte een breuk volgens het rapport van de oordelend radioloog. De mate van zichtbaarheid van de breuk werd beoordeeld als duidelijk, gemiddeld, en dubieus. Per lichaamsgebied was tenminste 25% van de breuken gemiddeld of dubieus. De 660 foto's werden gesplitst in tweemaal 330 foto's met vergelijkbare verdeling over lichaamsgebieden en zichtbaarheid van de breuk. →

Lezers
De 660 foto's werden voorgelegd aan twee radiologen, twee radiologen in opleiding, twee SEH-artsen, twee SEH-artsen in opleiding en twee radiologisch laboranten. Elke lezer las één set foto's.

Beoordelingen
De beoordeling van de foto's gebeurde in een gecontroleerde omgeving op de afdeling Radiologie, zodat alle lezers onder vergelijkbare omstandigheden oordeelden. De foto's werden tweemaal in een random volgorde aangeboden, eerst zonder AI-hulp, en na minimaal vier weken met hulp van AI. Bij de tweede presentatie werden de originele foto én de foto met de informatie van de GLEAMER AI naast elkaar aangeboden.
De lezers werd gevraagd aan te geven of ze een breuk zagen, hoeveel breuken er te zien waren en hoe zeker ze ervan waren dat er een breuk zat op een schaal van 1-5, waarbij 1 zeer onzeker en 5 zeer zeker was. De duur van de beoordeling van de serie werd genoteerd per sessie en per lezer.

Analyses
Diagnostische nauwkeurigheid, Area under Receiver Operating Curves (ROC) en sensitiviteit en specificiteit werden berekend als prestatiemetingen per lezer met en zonder AI-ondersteuning. De significantie van verschillen werd beoordeeld met behulp van een T-test.

Verschillen in prestaties tussen foto's van volwassenen en kinderen werden beschreven.

Prestatieverschillen tussen lezersgroepen werden bestudeerd met een post-hoc analyse, waarbij werd geëvalueerd hoe AI-assistentie de sensitiviteit en specificiteit beïnvloedde voor verschillende professionele achtergronden.

Statistische tests werden uitgevoerd met een significantie-niveau van $P < 0,05$. De analyses werden uitgevoerd met behulp van R-studio (versie 2023.03.0+386, Posit Software, PBC; the R Foundation for Statistical Computing).

Resultaten
De 660 foto's waren afkomstig van 658 patiënten. De gemiddelde leeftijd van de patiënten was 41,6 jaar (sd 22,8). Onder de patiënten waren 147 kinderen met een leeftijd tussen de 2 en 18 jaar (22,3%).

De verdeling van de foto's over lichaamsgebied en moeilijkheidsgraad staat in tabel 1.

Gouden standaard en beoordeling door AI
De gouden standaard werd bepaald door een dubbele beoordeling: een ervaren (MSK-)radioloog (MSK=Musculoskeletale) en een ervaren radioloog beoordeelden de aanwezigheid van fracturen met beschikbaarheid van follow-up onderzoeken en klinische context. In geval van onenigheid beoordeelde een tweede MSK-radioloog de informatie. Ten opzichte van de gouden standaard bereikte de GLEAMER AI-beoordeling een sensitiviteit van 96%, een specificiteit van 93% en een AUC van 0,94 voor de foto's van volwassenen. Het maakte daarbij niet uit van welk lichaamsgebied de foto's waren (tabel 2).

Bij de beoordeling van de röntgenfoto's van kinderen presteerde AI iets minder dan bij volwassenen, met een sensitiviteit van 98%, een specificiteit van 82% en een AUC van 0,9. Dit verschil met volwassenen was niet statistisch significant.

Er werden in totaal elf foto's fout-positief beoordeeld en vijf breuken gemist door AI. Daarbij kwamen fout-positieve beoordelingen voor bij alle lichaamsgebieden. De breuken werden gemist op foto's van elleboog (twee keer) en enkel/voet (drie keer).

Diagnostische beoordeling, vertrouwen en leessnelheid
Bij alle lezers gingen sensitiviteit, specificiteit en AUC significant omhoog. Gemiddeld steeg de sensitiviteit van 80% naar 90% ($p<0,01$). De specificiteit ging omhoog van 84% naar 91% ($P<0,01$). De AUC verbeterde van 0,82 naar 0,91 ($p<0,01$). Daarbij ging het vertrouwen in de diagnose van gemiddeld 4,45 naar gemiddeld 4,74 ($p<0,01$). De benodigde tijd voor het lezen nam gemiddeld met 21% af.

Tabel 1. Verdeling van röntgenfoto's over lichaamsgebied en moeilijkheidsgraad

Lichaamsgebied	Positief				Negatief
	Duidelijk	Gemiddeld	Dubieus	Totaal	
Schouder	32	6	9	47	51
Elleboog	28	11	7	46	58
Pols/hand	26	13	11	50	52
Bekken	28	15	7	50	51
Knie	31	11	5	47	54
Enkel/voet	19	19	14	52	52

Tabel 2. GLEAMER resultaat per lichaamsgebied

Lichaamsgebied	Sensitiviteit (%)	Specificiteit (%)	AUC
Schouder	98	86	0,92
Elleboog	96	86	0,91
Pols/hand	98	88	0,92
Bekken	98	96	0,98
Knie	98	94	0,96
Enkel/voet	91	90	0,91

Tabel 3. Resultaten per lezersgroep

	Zonder AI			Met AI		
Lezer	Sensitiviteit	Specificiteit	AUC	Sensitiviteit	Specificiteit	AUC
Radiologen	91,5%	81,6%	0,87	93,9%	90,0%	0,92
AIOS Radiologie	80,8%	84,0%	0,82	94,2%	90,1%	0,92
SEH Artsen	74,7%	87,0%	0,81	88,1%	90,7%	0,89
AIOS SEH	81,4%	82,2%	0,82	90,5%	92,2%	0,91
Laboranten	72,9%	84.9%	0,79	85,7%	93,4%	0,90

Subgroepanalyses

Lezers
Voor alle typen lezers stegen sensitiviteit, specificiteit en AUC bij gebruik van AI (tabel 3). Het effect leek het kleinst voor de ervaren radiologen en het grootst voor de artsen in opleiding en laboranten.Bij de laboranten was de kwaliteit van de beoordeling aanvankelijk het meest variabel, wat door de toevoeging van AI sterk verminderde. Met behulp van AI bereikte de kwaliteit van de beoordeling van alle lezersgroepen bijna het niveau van de ervaren radiologen. De sensitiviteit bleef iets achter voor laboranten en SEH-artsen. De zekerheid over de diagnose steeg met name bij de arts-assistenten, zowel van de Radiologie als van de SEH.

Voor alle lezers was de leestijd significant korter met AI, behalve voor de laboranten. De mate van verkorting varieerde van 7% voor de radiologen tot 39% voor de SEH-artsen.
De laboranten hadden beiden meer tijd nodig mét AI dan zonder.

Duidelijkheid breuk
De sensitiviteit, specificiteit en AUC hingen samen met de duidelijkheid van de breuk. De duidelijke breuken werden beter gevonden dan de gemiddelde breuken en deze werden weer beter gediagnosticeerd dan de dubieuze breuken.
Met het gebruik van AI nam de sensitiviteit toe met 3% voor de duidelijke breuken, met 15% voor de gemiddelde breuken en met 24% voor de dubieuze breuken ($p<0.01$).

Discussie en conclusie

De resultaten van deze studie laten zien dat het gebruik van AI voor het detecteren van breuken van meerwaarde kan zijn in de spoedzorg. Het diagnostisch oordeel van AI als standalone kwam opvallend vaak overeen met de gouden standaard. GLEAMER deed daarbij niet onder voor de radiologen in de studie. Dat zou kunnen betekenen dat niet-radiologische specialisten, zoals de huisarts, (mede) op basis van het oordeel van de GLEAMER AI sneller zouden kunnen besluiten of patiënten moeten worden doorverwezen naar de SEH. Dit scheelt potentieel veel wachttijd en doorverwijzingen. De goede standalone performance van Gleamer geeft ook laboranten de mogelijkheid patiënten direct naar de SEH door te verwijzen, zonder voor iedere patiënt op het oordeel van de spoedradioloog te moeten wachten. Dubieuze Gleamer-beoordelingen zouden wel met spoed door de radioloog beoordeeld moeten worden. Uiteraard worden alle beelden uiteindelijk door een radioloog beoordeeld, zodat eventuele misinterpretaties gecorrigeerd kunnen worden.

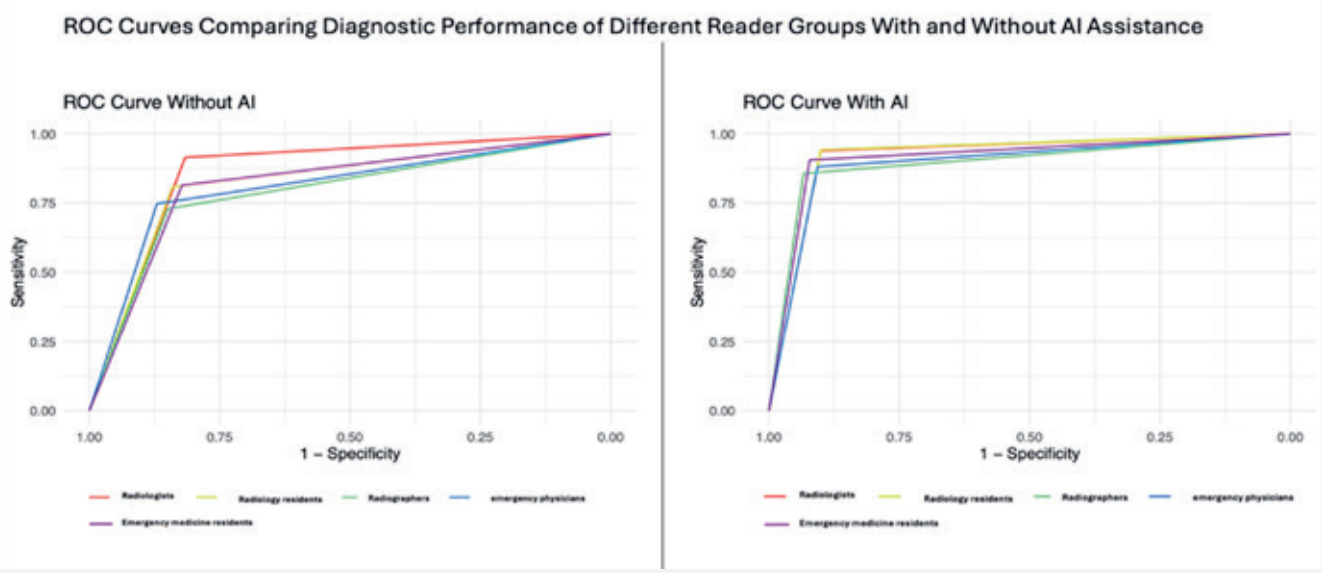
Er is natuurlijk al eerder onderzoek gedaan naar het functioneren van de GLEAMER AI. Dit eerdere onderzoek in andere landen met weliswaar andere apparatuur liet vergelijkbare resultaten zien, met een hogere sensitiviteit en specificiteit voor verschillende lichaamsgebieden, voor zowel volwassenen als kinderen bij gebruik van GLEAMER.

Ook onze bevinding dat verschillende beroepsgroepen beter oordelen bij gebruik van AI is eerder gepubliceerd. Dat de toename van de sensitiviteit en specificiteit bij moeilijk te beoordelen foto's het grootst is, werd al eerder gesuggereerd door Umapathy et al.

In deze studie hadden de lezers niet de beschikking over klinische context en patiëntinformatie. Dat kan ervoor hebben gezorgd dat met name de SEH-artsen lager scoorden, omdat zij normaliter deze informatie betrekken in hun diagnostiek. GLEAMER AI gebruikt evenmin klinische data in het oordeel. Mogelijk wordt het oordeel van AI nog beter indien deze informatie kan worden toegevoegd. Een tweede kritiekpunt op het onderzoek kan zijn dat het tweemaal laten lezen van dezelfde foto's mogelijk tot herinnering van de beoordeling kan hebben geleid. Hoewel er tenminste vier weken zat tussen beide sessies, is dit mogelijk. Het is echter niet waarschijnlijk dat dit de verbetering in de diagnostiek verklaart.

In totaal werden er tien lezers gevraagd, twee per lezersgroep. Dit maakt de uiteindelijke conclusies per lezersgroep minder precies, al was de verbetering van fractuurdetectie in alle subgroepen zichtbaar en statistisch significant.

Onze resultaten suggereren dat de invoering van de GLEAMER AI op de SEH en de Huisartsenpost meerwaarde zou kunnen hebben. Of ook de kosten dalen en patiënten zowel goedkoper uit zijn als sneller thuis, moet nog worden onderzocht. ●



Figuur 1. ROC curves van de oordelen zonder en met AI van de verschillende beroepsgroepen

Sensitiviteit, specificiteit en ROC curves

Als je wilt vaststellen of een test helpt bij diagnostiek heb je een aantal mogelijkheden. Als de test een ja/nee antwoord geeft, bijvoorbeeld knik in de arm op een andere plek dan de elleboog of pols, kun je de sensitiviteit en specificiteit berekenen. Deze geven respectievelijk aan hoeveel procent van de zieken als "ziek" uit de test komen (sensitiviteit), en hoeveel procent van de niet-zieken als "niet-ziek" uit de test komen (specificiteit).
Maar als je een continue maat hebt kun je geen twee-bij-twee tabel tekenen en dus ook niet eenvoudig de sensitiviteit en specificiteit uitrekenen. Dan moet je een afkappunt vaststellen en op basis daarvan mensen als test-positief ("ziek") en test-negatief ("niet-ziek") indelen zodat je weer een sensitiviteit en specificiteit kunt uitrekenen. Maar wat zouden de sensitiviteit en de specificiteit zijn als het afkappunt ergens anders zou liggen? Dat kun je ook uitrekenen.
Als je voor alle afkappunten de sensitiviteit en specificiteit hebt uitgerekend kun je op basis daarvan een ROC curve tekenen. Op de x-as staat 1-specificiteit en op de y-as staat de sensitiviteit.
Als de curve helemaal links boven in de hoek komt heb je een perfecte test. Hoe meer de curve naar het midden zakt, hoe slechter de test. Dat druk je uit in de AUC, de Area Under the Curve. Links in de hoek is deze AUC 1, en als de lijn door het midden loopt is de AUC 0,5. Je kunt daarmee dus in een plaatje voor verschillende testen laten zien hoe goed ze zijn. In dit artikel is de ROC curve gebaseerd op de (enige) sensitiviteit en specificiteit die per groep kon worden berekend.