Artificiële intelligentie in de huisartsenpraktijk

Olivier Nauwynck, Universiteit Gent

Promotor: Prof. dr. Stefan Heytens, Universiteit Gent
Co-promotor: dr. Veerle Piessens, Universiteit Gent

Master of Family Medicine
Masterproef Huisartsgeneeskunde
Academiejaar: 2023 – 2024

Deze masterproef is een examendocument dat niet werd gecorrigeerd voor eventueel vastgestelde fouten. Zonder voorafgaande schriftelijke toestemming van zowel de promotor(en) als de auteur(s) is overnemen, kopiëren, gebruiken of realiseren van deze uitgave of gedeelten ervan verboden. Voor aanvragen tot of informatie i.v.m. het overnemen en/of gebruik en/of realisatie van gedeelten uit deze publicatie, wendt u tot de universiteit waaraan de auteur is ingeschreven.
Voorafgaande schriftelijke toestemming van de promotor(en) is eveneens vereist voor het aanwenden van de in dit afstudeerwerk beschreven (originele) methoden, producten, schakelingen en programma's voor industrieel of commercieel nut en voor de inzending van deze publicatie ter deelname aan wetenschappelijke prijzen of wedstrijden.

Voorwoord

Via deze weg wil ik enkele mensen bedanken. Ten eerste wil ik mijn (co-) promotoren Prof. Dr. Stefan Heytens en Dr. Veerle Piessens bedanken om mij de kans te geven om dit onderwerp te mogen uitwerken en voor de constructieve feedbackmomenten. Daarnaast wil ik ook mijn vriendin en familie bedanken voor de aanmoediging en steun. Dit niet alleen tijdens het schrijven van deze masterproef, maar ook tijdens de afgelopen jaren.

Abstract

Context. Het takenpakket van een huisarts is uitgebreid en huisartsen staan vaak voor complexe casussen. Het is echter geen geheim dat dit werk onderhevig is aan menselijke fouten en de kwaliteit van de zorg verschilt. De steeds ouder wordende patiëntenpopulatie die gepaard gaat met meer complexe ziektebeelden zorgt ervoor dat deze uitdagingen alleen maar groter worden. Artificiële intelligentie (AI) zou hulp kunnen bieden bij het aanpakken van deze uitdagingen. Ook zouden er winsten kunnen geboekt worden op zowel economisch vlak als op het vlak van gezondheid. Dit wordt in de hand gewerkt door de steeds toenemende hoeveelheid medische data, gepaard met steeds krachtiger wordende AI-algoritmes. Hoewel het aantal studies exponentieel toegenomen is sinds 2015, is er sinds de review van Abbasgholizadeh Rahimi et al. (1), geen systematische literatuurstudie gedaan om zicht te krijgen op de huidige stand van zaken van AI-toepassingen binnen de huisartsgeneeskunde. Dit is echter noodzakelijk om de tekortkomingen van de huidige studies te belichten, want dit kan op zijn beurt meer richting geven aan toekomstig onderzoek. De huidige studie had dan ook tot doel om een algemeen overzicht te geven met betrekking tot het eerdere onderzoek over AI-toepassingen in de huisartsenpraktijk.
Onderzoeksvraag. Aan de hand van een scoping review werd de onderzoeksvraag, "Welke toepassingen van AI in de huisartsenpraktijk zijn er onderzocht en wat zijn de resultaten?", in kaart gebracht.
Methode. Om deze scoping review systematisch uit te voeren werd een stappenplan gevolgd, gebaseerd op beschreven richtlijnen. Aan de hand van een initiële verkennende literatuurstudie werd de onderzoeksvraag verfijnd, werden in- en exclusiecriteria opgesteld en werd een zoekstrategie via Pubmed, Web of Science en Scopus uitgewerkt.
Resultaten. Er werden 1132 studies geïdentificeerd na verwijdering van duplicaten. Na een systematisch selectieproces werden 68 studies, gepubliceerd tussen januari 2018 en november 2023, geïncludeerd. Het aantal studies steeg exponentieel tussen 2018 en 2023. 41% van de studies had als doel om AI-toepassingen te ontwikkelen. 24% van de studies beoogde om de AI-toepassingen te valideren en 35% van de studies deed een combinatie van beide. Hoewel de helft van de studies als doel had om reeds ontwikkelde AI-modellen te valideren in een nieuwe dataset, waren er slechts tien studies (15%) die AI-toepassingen prospectief onderzochten. Verder beoogde de meerderheid (72%) van de AI-toepassingen om diagnostiek te ondersteunen. Daarnaast kenden de onderzochte AI-modellen vooral hun toepassing binnen cardiologie (21%), dermatologie (16%) en geriatrie (10%). Een grote variabiliteit werd opgemerkt in de manier waarop methodologie en resultaten gerapporteerd werden.
Conclusie. Hoewel de meeste geïncludeerde studies veelbelovende AI-toepassingen onderzochten, zijn er nog meer stappen nodig om tot een effectieve implementatie te komen. Deze scoping review maakt duidelijk dat het huidige onderzoek naar AI-toepassingen nog in zijn kinderschoenen staat. Meer implementatieonderzoek is noodzakelijk om de toepasbaarheid, accuraatheid en meerwaarde van deze toepassingen te onderzoeken, alvorens ze kunnen gebruikt worden in de praktijk. Zes van de tien prospectieve validatie studies werden uitgevoerd in 2022-2023. Dit toont een positieve trend aan in het onderzoeksveld.

Inhoudstafel

Inleiding..... 1
Artificiële intelligentie..... 1
Artificiële intelligentie in de huisartsgeneeskunde..... 3
Huidige studie..... 4
Methode..... 5
Stap 1: verfijnen van de onderzoeksvraag..... 5
Stap 2: in- en exclusiecriteria bepalen..... 5
Stap 3: formuleren van relevante zoektermen..... 6
Stap 4: identificeren van relevante databanken..... 6
Stap 5: uitwerken van zoekstrategie..... 6
Stap 6: studieselectie door middel van screening in twee fases..... 8
Stap 7: data-extractie en synthese..... 8
Stap 8: interpretatie en rapportage van de resultaten..... 10
Resultaten..... 11
Overzicht van resultaten..... 11
Publicatiejaar van de studies..... 11
Doel van de studies..... 11
Type artificiële intelligentie..... 12
Functie artificiële intelligentie..... 13
Subdomein geneeskunde..... 13
Discussie..... 45
De toepassing van AI binnen de huisartsgeneeskunde: de huidige stand van zaken..... 45
Kritische reflectie en implicaties voor toekomstig onderzoek..... 45
Limitaties..... 49
Conclusie..... 50

Inleiding

Artificiële intelligentie (AI) is in de afgelopen decennia uitgegroeid tot een baanbrekende technologische revolutie die verschillende sectoren van de samenleving heeft getransformeerd. In de eerstelijnsgezondheidszorg, en specifiek in de huisartsenpraktijk, heeft AI het potentieel om aanzienlijke veranderingen teweeg te brengen. Toepassingen van AI in huisartsgeneeskunde worden momenteel actief ontwikkeld en onderzocht. Het doel van deze literatuurstudie is om een overzicht te geven van deze studies.

Artificiële intelligentie

Artificial intelligence, artificiële of kunstmatige intelligentie in het Nederlands, is het onderzoeksveld dat zich bezighoudt met intelligente systemen. Dit omvat zowel het doorgronden als het construeren ervan. Voor het laatste punt is het essentieel dat deze systemen op een doeltreffende en veilige manier een bepaald gewenst gedrag vertonen en dat in variabele omstandigheden (2).

Het is belangrijk om aan te stippen dat het onderscheid vaak niet duidelijk is tussen AI, *machine learning* en *deep learning*. Vooral in die laatste twee zijn er recentelijk grote successen geboekt. *Machine learning*, zelf een soort AI, is de koepelterm voor systemen die data gebruiken om een bepaalde taak te leren uitvoeren en hierbij beter worden met ervaring. Deze systemen zijn flexibel en kunnen naargelang de aangeleverde data en het leeralgoritme verschillende taken leren uitvoeren. *Deep learning* is een specifieke vorm van *machine learning* die vooral het afgelopen decennium een (her-)opmars beleefde. Bij *deep learning* worden neurale netwerken gebruikt. Neurale netwerken bestaan uit geneste lagen van neuronen, computationele eenheden die bepaalde inputwaarden transformeren tot een output. Het zijn simpele non-lineaire functies waar de parameters kunnen worden geleerd. Het geheel van deze eenvoudige, leerbare functies, diep en breed genest, laat toe dat neurale netwerken als geheel complexe functies kunnen voorstellen (2,3).

Om AI-modellen te ontwikkelen worden datasets verdeeld in verschillende subsets (3,4): een trainingset, een validatieset en een testset. De trainingset maakt gebruik van het grootste deel van de dataset, en wordt gebruikt om AI-modellen te trainen. De data van deze trainingset vormt dus de basis waarop de parameters van het AI-model zich ontwikkelen. Vervolgens wordt het AI-model toegepast op de validatieset, om het AI-model te testen op een nieuw deel van de dataset, en vervolgens nog te optimaliseren. De hyperparameters kunnen in deze fase worden aangepast, zodat het model dat ontwikkeld werd op de trainingset zo optimaal mogelijk presteert op de validatieset. Tot slot kan het getrainde AI-model worden toegepast op een testset om te onderzoeken hoe goed het AI-model presteert op ongeziene data en dus in welke mate het model generaliseerbaar is.

De geschiedenis van AI strekt over tientallen jaren en wordt gekenmerkt door pieken en dalen in behaalde successen en interesse in het veld. Een exacte startdatum is moeilijk vast te pinnen. Zo kan men de start situeren bij de wiskundige regels die Leibniz (1676) en Cauchy (1847) afleidden en cruciaal bleken voor lerende algoritmes (5). Echter is het een

conventie om de geschiedenis van de moderne AI te starten in de jaren '40 van de afgelopen eeuw (2).

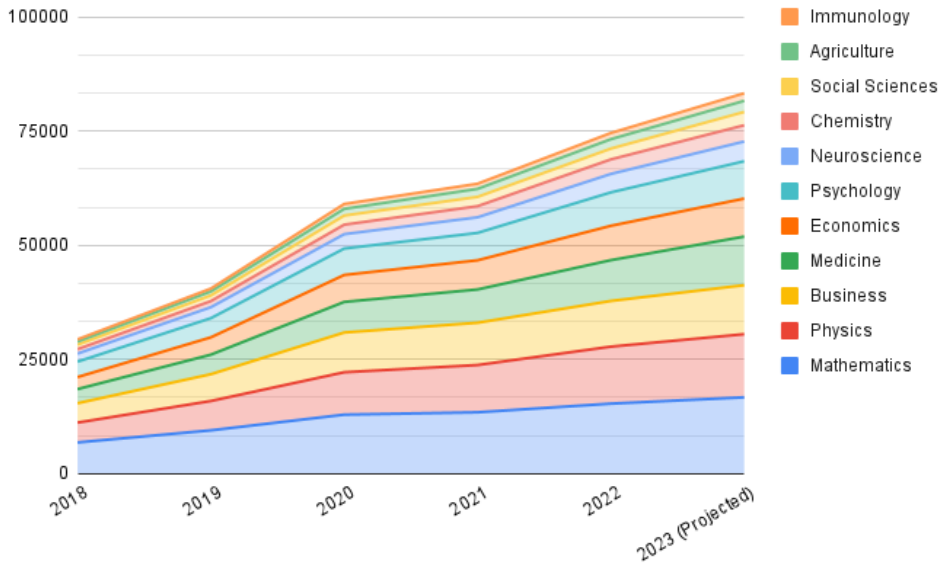
In wat volgt kan men verschillende periodes onderscheiden (2). De jaren tussen 1943 en 1956 en tussen 1952 en 1969 zijn respectievelijk de periodes waarin de fundamenten van AI gelegd werden en er eerste successen geboekt werden. In de laatste van deze periodes ontstond het vakgebied cybernetica, een vroege voorloper van het hedendaagse *deep learning*. Zo lieten algoritmes zoals de perceptron (6) of ADALINE (7) toe dat de parameters van een neuron geleerd werden. Wat volgde was de eerste 'AI winter' (1966-1973). Ondanks het grote optimisme dat eerdere periodes kenmerkte, bleef het vermogen om complexe taken op te lossen uit en werden verwachtingen niet ingelost. Dit zorgde voor een gedaalde interesse in verder onderzoek .

De interesse herleefde tussen 1969 en 1986, voornamelijk door de opkomst van *expert systems*. In deze systemen kan men domeinkennis inbrengen, waarna aan de hand van regels en zich beroepende op deze kennis bepaalde taken volbracht kunnen worden. Een primeur voor de medische wereld was MYCIN (8) dat bepaalde bacteriën kon detecteren met nagenoeg dezelfde accuraatheid als specialisten. Voor complexe problemen bleken *expert systems* niet krachtig genoeg en een nieuwe AI winter volgde (1986-2000). Toch werden in deze periode belangrijke doorbraken gerealiseerd. Het connectionisme, met de heruitvinding van de '*backpropagation rule*' was essentieel voor het trainen van diepe neurale netwerken. Probabilistische modellen kenden ook hun doorbraak. Data, statische modellering, optimalisatietechnieken en *machine learning* werden samengebracht. Zo ontstond er een synergie tussen onderzoeksgebieden, zoals robotica, *computer vision* en spraakherkenning, die voordien eerder onafhankelijk opereerden. De kiem was gelegd voor de volgende succesvolle periodes (3).

Cruciaal voor het recente succes was het massale gebruik van het wereldwijde web en de vele data die het opleverde. Dit wordt ook wel *big data* genoemd. Het zorgde ervoor dat *machine learning* steeds betere resultaten gaf. Met geoptimaliseerde hardware, zoals GPU's of TPU's, is het ook mogelijk geworden om steeds grotere neurale netwerken te laten leren. Sinds de doorbraak in beeldherkenning in de ImageNet competitie (2011) (9) ging het in een sneltempo vooruit. Voorbeelden van bekende mijlpalen zijn AlphaGo (10), BERT (11), DALL-E (12) en de GPT-1/-2/-3 modellen (13–15). Talloze architecturen voor neurale netwerken werden ontwikkeld, met als enkele voorbeelden Generative Adversarial Networks (GANs) (16), Neural Radiance Fields (NeRFs) (17), Denoising Diffusion Models (18) en Transformers (19). Een recentere trend is het combineren van modaliteiten zoals bijvoorbeeld beeld en tekst. Dit gebeurde ook voor medische toepassingen, zoals bijvoorbeeld in Med-PaLM M (20). In november 2022 won AI, met de lancering van ChatGPT (21), dat als onderliggend AI model GPT-3, -3.5 of -4 gebruikt, op grote schaal de aandacht van de buitenwereld.

AI is dus de afgelopen decennia uitgegroeid tot een baanbrekende technologische revolutie die verschillende sectoren van de samenleving heeft getransformeerd (2). Ook binnen de gezondheidszorg is er volop gedigitaliseerd. Zo maken artsen gebruik van digitale patiëntendossiers en zijn er vele soorten medische data beschikbaar. Denk hierbij bijvoorbeeld aan medische beelden, patiëntendossiers en genetische data. De digitalisering laat toe om AI-toepassingen te ontwikkelen en gebruiken (22). Figuur 1, overgenomen uit het recente rapport ‘State of AI 2023’ (23), illustreert de interesse in AI-toepassingen voor

geneeskunde. Het toont dat uit twintig verschillende wetenschappelijke disciplines het aantal AI-publicaties voor geneeskunde het sterkst stijgt.



Figuur 1. Overzicht AI-publicaties in verschillende disciplines (23)

Artificiële intelligentie in de huisartsgeneeskunde

De huisartsenpraktijk is een essentieel onderdeel van de gezondheidszorg. Het takenpakket van een huisarts is uitgebreid, variërend van preventieve zorg en diagnose tot behandeling en verwijzing naar specialisten. Huisartsen staan vaak voor complexe casussen, waarbij ze moeten vertrouwen op hun medische kennis en klinische ervaring om de best mogelijke zorg te bieden. Het is echter geen geheim dat deze processen vaak onderhevig zijn aan menselijke fouten en de kwaliteit van de zorg verschilt (24). De steeds ouder wordende patiëntenpopulatie die gepaard gaat met meer complexe ziektebeelden zorgt ervoor dat deze uitdagingen alleen maar groter worden (25).

Deze uitdagingen zouden eventueel deels kunnen aangepakt worden door AI te gebruiken. Meer bepaald kunnen geavanceerde AI-gebaseerde computerprogramma's en algoritmen worden ontwikkeld die in staat zijn om enorme hoeveelheden medische gegevens te analyseren, patronen te ontdekken en het klinisch redeneren te ondersteunen. AI kan het werk van de arts aanvullen en verlichten, waardoor de behoeften van de individuele patiënt beter kunnen worden afgestemd op de zorg die zij ontvangen (26,27). Ook vanuit een gezondheidseconomisch perspectief kan het gebruik van AI de kosten verlagen (28).

Hoewel AI potentieel heeft om uitdagingen binnen de huisartsgeneeskunde aan te pakken, zijn er ook een aantal nieuwe uitdagingen die hand in hand gaan met het gebruik van AI. Op ethisch vlak moet er bijvoorbeeld rekening gehouden worden met privacy, verminderde accuraatheid van AI-modellen in minderheidsgroepen (bias en discriminatie), transparantie (inzicht in de besluitvorming van AI-modellen) en toegang tot de algoritmen, de gevolgen voor de arts-patiënt relatie, verantwoordelijkheid, autonomie etc (29). Verder worden er ook moeilijkheden gerapporteerd bij de implementatie van AI-toepassingen, zoals

bijvoorbeeld kosten en terugbetaling van AI-diagnostiek of -screening (30), een gebrek aan technische infrastructuur (31) en de nood aan opleidingen over AI-toepassingen (32). Daarnaast heerst er bezorgdheid over de mogelijke vervanging van menselijke zorgverleners door AI (33,34). Om AI-toepassingen efficiënt te kunnen gebruiken in de huisartsenpraktijk, is het ook nodig om deze eerst uitgebreid te onderzoeken. De voorbije jaren is het onderzoek naar medische toepassingen van AI toegenomen, resulterend in verschillende reviews over bijvoorbeeld het gebruik van AI in dermatologie (35), cardiologie (36), radiografie (37) en in de algemene eerstelijnszorg (38,39). Daarnaast zijn er talrijke studies over AI-toepassingen in de huisartsenpraktijk. Hoewel het aantal studies exponentieel blijft toenemen sinds 2015, is er sinds de review van Abbasgholizadeh Rahimi et al. (1), die studies includeerde tot februari 2020, geen systematische literatuurstudie gedaan om zicht te krijgen op de huidige stand van zaken van AI-toepassingen binnen de huisartsgeneeskunde. Dit is echter noodzakelijk om de tekorten van de huidige studies te belichten, want dit kan op zijn beurt meer richting geven aan toekomstig onderzoek. Ook voor huisartsen die werken in de klinische praktijk is een huidige stand van zaken met betrekking tot eerder onderzoek over AI-toepassingen nodig.

Huidige studie

De huidige studie heeft dan ook tot doel om een algemeen overzicht te geven met betrekking tot het eerdere onderzoek over AI-toepassingen in de huisartsenpraktijk. Aan de hand van een scoping review wordt de onderzoeksvraag, "Welke toepassingen van AI in de huisartsenpraktijk zijn er onderzocht in en wat zijn de resultaten?”, uitgebreid in kaart gebracht. Een scoping review is een geschikte methode om een beter beeld te krijgen over de huidige stand van zaken in onderzoek. Een scoping review laat ons toe om deze onderzoeksvraag breed in kaart te brengen. Dit kan uiteindelijk richting geven aan toekomstig onderzoek (40).

Methode

Om deze scoping review systematisch uit te voeren werd een stappenplan gevolgd gebaseerd op Kazi et al. (40). Aan de hand van een initiële verkennende literatuurstudie werd de onderzoeksvraag verfijnd, werden in- en exclusiecriteria opgesteld en een zoekstrategie uitgewerkt. Vervolgens werden de studies geselecteerd in twee fasen (zie ook Figuur 2). Onderstaand worden de verschillende stappen in detail besproken.

Stap 1: verfijnen van de onderzoeksvraag

Aan de hand van de PICO-methode werd de onderzoeksvraag verfijnd. In deze scoping review werd de onderzoeksvraag “Welke toepassingen van artificiële intelligentie zijn er onderzocht in de huisartsenpraktijk en wat zijn de resultaten?” bestudeerd.

P: (patiënten van) huisartsen

I: toepassingen van artificiële intelligentie

C:/

O: ontwikkeling en/of validatie van AI-toepassingen

Stap 2: in- en exclusiecriteria bepalen

De in- en exclusiecriteria worden weergegeven in Tabel 1.

Tabel 1: in- en exclusiecriteria

Inclusiecriteria	**Exclusiecriteria**
(Patiënten van) huisartsen zijn (minstens een deel van) de steekproef. Er zijn geen specifieke leeftijds-, geslachts- of demografische beperkingen voor deze populatie.	Studies die zich uitsluitend richten op AI in andere medische specialismen (bijvoorbeeld radiologie, pathologie en cardiologie). Studies die handelen over andere zorgberoepen binnen de eerstelijnszorg zonder rechtstreeks verband met huisarts.
Studies die een toepassing van AI onderzoeken. Studies die betrekking hebben op het gebruik van artificiële intelligentie (AI) technologieën, zoals machine learning, deep learning, natural language processing, expertsystemen of andere AI-technieken. Studies over AI-toepassingen voor elk aspect van huisartsgeneeskunde, waaronder maar niet beperkt tot diagnose, behandeling, patiëntenbeheer, risicobeoordeling en inschatting.	Studies waarbij AI gebruikt wordt als methodologie om een specifieke onderzoeksvraag te beantwoorden (bv. AI wordt gebruikt om predictoren te identificeren).

Studies gepubliceerd tussen januari 2018 en november 2023.	
Kwalitatieve en kwantitatieve studies.	Reviews, preregistraties en opiniestukken.
Studies geschreven in het Engels, Frans of het Nederlands.	Studies geschreven in een andere taal.

Stap 3: formuleren van relevante zoektermen

Op basis van een initiëel verkennende literatuurstudie en 'MeSH' termen werden specifieke zoektermen voor verschillende concepten geïdentificeerd (zie ook Tabel 2).

Tabel 2: concepten en bijhorende zoektermen

Concept	**Zoektermen**
Artificiële intelligentie	Intelligence, Artificial; Computational Intelligence; Intelligence, Computational; Machine Intelligence; Machine learning; Learning, machine; Deep learning; Hierarchical learning; Supervised machine learning; Active machine learning; Computer Reasoning; AI (Artificial Intelligence) Computer Vision Systems; Knowledge Acquisition; Knowledge Representation; Chatbot; Large language model Deep learning;Neural networks
Huisartsgeneeskunde	Family medicine; Family practice; General practice; Primary care; Primary care, physician;;General practitioner

Stap 4: identificeren van relevante databanken

Op basis van een initiëel verkennende literatuurstudie werden de meest relevante databanken geïdentificeerd: Pubmed, Scopus en Web Of Science.

Stap 5: uitwerken van zoekstrategie

De verschillende zoekstrategieën per databank worden weergegeven in Tabel 3.

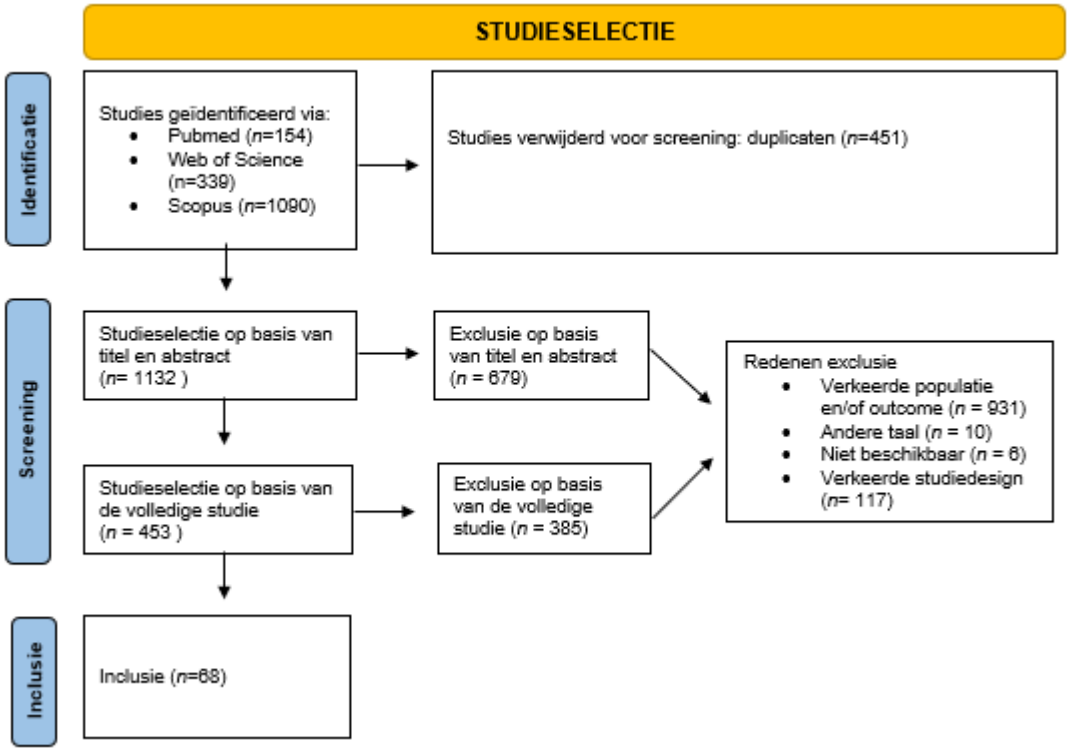
Tabel 3: zoekstrategie per databank

Databank	Zoekstrategie
Pubmed	("Family Practice"[Mesh] OR "General Practice"[Mesh] OR "Primary Health Care"[Mesh] OR "Family Medicine"[Mesh] OR "General Practitioners"[Mesh] OR "Physicians, Family"[Mesh]) OR ('Family Practice'[Title/Abstract] OR 'General Practice'[Title/Abstract] OR 'Primary Health Care'[Title/Abstract] OR 'Family Medicine'[Title/Abstract] OR 'General Practitioners'[Title/Abstract] OR 'Physicians, Family'[Title/Abstract] OR 'Primary Care Physician'[Title/Abstract]) AND ("Artificial Intelligence"[Mesh] OR "Computational Intelligence"[Mesh] OR "Machine Learning"[Mesh] OR "Deep Learning"[Mesh] OR "Supervised Machine Learning"[Mesh] OR "Computer Vision"[Mesh] OR "Knowledge Acquisition"[Mesh] OR "Knowledge Representation"[Mesh] OR "Chatbots"[Mesh] OR "Large Language Model"[Mesh] OR "Neural Networks"[Mesh]) AND ('Family Practice'[Title/Abstract] OR 'General Practice'[Title/Abstract] OR 'Primary Health Care'[Title/Abstract] OR 'Family Medicine'[Title/Abstract] OR 'General Practitioners'[Title/Abstract] OR 'Physicians, Family'[Title/Abstract] OR 'Primary Care Physician'[Title/Abstract]) AND ('Artificial Intelligence'[Title/Abstract] OR 'Computational Intelligence'[Title/Abstract] OR 'Machine Learning'[Title/Abstract] OR 'Deep Learning'[Title/Abstract] OR 'Supervised Machine Learning'[Title/Abstract] OR 'Computer Vision'[Title/Abstract] OR 'Knowledge Acquisition'[Title/Abstract] OR 'Knowledge Representation'[Title/Abstract] OR 'Chatbots'[Title/Abstract] OR 'Large Language Model'[Title/Abstract] OR 'Neural Networks'[Title/Abstract])
Web of Science	TS=("Family Practice" OR "General Practice" OR "Primary Health Care" OR "Family Medicine" OR "General Practitioners" OR "Physicians, Family" OR "Primary Care Physician") AND TS=("Artificial Intelligence" OR "Computational Intelligence" OR "Machine Learning" OR "Deep Learning" OR "Supervised Machine Learning" OR "Computer Vision" OR "Knowledge Acquisition" OR "Knowledge Representation" OR "Chatbots" OR "Large Language Model" OR "Neural Networks")
Scopus	(TITLE-ABS-KEY("Family Practice") OR TITLE-ABS-KEY("General Practice") OR TITLE-ABS-KEY("Primary Health Care") OR TITLE-ABS-KEY("Family Medicine") OR TITLE-ABS-KEY("General Practitioners") OR TITLE-ABS-KEY("Physicians, Family") OR TITLE-ABS-KEY("Primary Care Physician")) AND (TITLE-ABS-KEY("Artificial Intelligence") OR TITLE-ABS-KEY("Computational Intelligence") OR TITLE-ABS-KEY("Machine Learning") OR TITLE-ABS-KEY("Deep Learning") OR TITLE-ABS-KEY("Supervised Machine Learning") OR TITLE-ABS-KEY("Computer Vision") OR TITLE-ABS-KEY("Knowledge Acquisition") OR TITLE-ABS-KEY("Knowledge Representation") OR TITLE-ABS-KEY("Chatbots") OR TITLE-ABS-KEY("Large Language Model") OR TITLE-ABS-KEY("Neural Networks") OR TITLE-ABS-KEY("Intelligence, Artificial") OR TITLE-ABS-KEY("Computational Intelligence") OR TITLE-ABS-KEY("Machine Intelligence") OR TITLE-ABS-KEY("Machine learning") OR TITLE-ABS-KEY("Learning, machine") OR TITLE-ABS-KEY("Deep learning") OR TITLE-ABS-KEY("Hierarchical learning") OR TITLE-ABS-KEY("Supervised machine learning") OR TITLE-ABS-KEY("Active machine learning") OR TITLE-ABS-KEY("Computer Reasoning") OR TITLE-ABS-KEY("AI (Artificial Intelligence)") OR TITLE-ABS-KEY("Computer Vision Systems"))

Stap 6: studieselectie door middel van screening in twee fases

Hoewel het aangeraden is om de studieselectie door twee onafhankelijke beoordelaars te laten gebeuren, was dit niet mogelijk in het kader van deze masterproef. Eén beoordelaar heeft de verschillende studies geselecteerd.

In de eerste fase werden studies geselecteerd op basis van titels en abstracts. In de tweede fase werden de overgebleven studies volledig gelezen, om zo tot een finale beslissing te komen. In Figuur 2 worden deze verschillende fasen en de redenen voor exclusie weergegeven.



Figuur 2. Flowchart studieselectie.

Stap 7: data-extractie en synthese

Gelijkaardig aan de data extractie in de scoping review van Keuper et al. (38) werden onderstaande gegevens uit de geïncludeerde studies verzameld (zie ook Tabel 4).

Tabel 4: data-extractie

Basisinformatie	● Auteurs ● Titel ● Jaar van publicatie
Doel van de studie	Eén studie kan één of meerdere doelen hebben: ● Ontwikkeling (O): Het ontwikkelen of vernieuwen van een AI-toepassing. ● Validatie (V): Het testen van een AI-toepassing in een nieuwe dataset. Een ander voorbeeld is onderzoek waarbij de vergelijking wordt gemaakt tussen de resultaten van AI-modellen en de resultaten van artsen. ● Ander: het doel van de studie past niet binnen bovenstaande categorieën.
Type AI	Eén studie kan één of meerdere types AI bespreken: ● *Computer vision* (CV) is het deelgebied van AI dat algoritmes inzet om beeldmateriaal te verwerken en analyseren. De algoritmes kunnen bepaalde patronen in beelden herkennen en deze gebruiken om bepaalde inzichten te verkrijgen, zoals bijvoorbeeld de diagnose van kanker stellen aan de hand van een scan. ● *Data mining* (DM) wordt gebruikt om verbanden te vinden in grote hoeveelheden (ruwe) data. Hiervoor worden statische of machine learning-gebaseerde methoden gebruikt, zoals bijvoorbeeld het gebruik van data mining om inzichten te verkrijgen uit een grote hoeveelheid elektronische patiëntendossiers. Deze informatie kan bijvoorbeeld gebruikt worden om betere diagnoses te stellen of behandelingen uit te stippelen. ● *Natural language processing* (NLP) omvat het deelgebied van AI dat omgaat met het verwerken en analyseren van taal. Machine learning wordt gebruikt om uit taal, geschreven of gesproken, inzichten te verwerven, zoals bijvoorbeeld met *Large Language Models.* Deze zijn een vorm van diepe neurale netwerken, die een groot aantal parameters hebben. Natural language processing kan in een medische context gebruikt worden om bijvoorbeeld patiëntendossiers te analyseren. ● *Machine learning* (ML) is de verzameling van zelflerende systemen. Een algoritme kan aangeleerd worden om een bepaalde taak te voltooien. Doordat geannoteerde data (*labeled data* in het Engels) beschikbaar is, kan er tijdens het leerproces feedback gegeven worden aan het algoritme over de juistheid van het resultaat. Een algemeen voorbeeld van een toepassing is het voorspellen van een bepaalde diagnose gegeven bepaalde klinische informatie. Tijdens het leerproces kan de voorspelling van het algoritme getoetst worden aan de gekende diagnose. Deze feedback wordt vervolgens gebruikt door het zelflerend algoritme om zijn voorspelling te corrigeren. Soms wordt er geen gebruikgemaakt van geannoteerde data. De algoritmes worden aangewend om op een andere manier inzichten te halen uit bepaalde data. Het verschil zit in de feedback die gegeven wordt aan het algoritme, want hierbij is het niet nodig om bijvoorbeeld een gekende diagnose te hebben. Een voorbeeld hiervan is het clusteren van data in bepaalde subgroepen die verwant zijn. ● Ander*:* Het type past niet binnen bovenstaande categorieën.

Functie AI	Eén AI-toepassing kan één of meerdere functies hebben: ● Diagnostiek: AI geeft informatie over de kans dat een bepaalde patiënt een bepaalde diagnose heeft. ● Behandeling: AI geeft informatie om te helpen bij het kiezen van een behandeling bij een patiënt met bepaalde symptomen of diagnose(s). ● Doorverwijzing: AI geeft informatie over welke patiënten moeten doorverwezen worden naar specialisten, of AI helpt met het technische aspect van het doorverwijzing traject. ● Voorspelling: AI helpt om zaken in de toekomst te voorspellen, zoals bijvoorbeeld prognose en de noodzaak van een spoedopname. ● Informatie extractie: AI wordt gebruikt om kennis te verzamelen uit datasets voor verder gebruik. ● Andere: het doel van de toepassing
Subdomein binnen geneeskunde	Het subdomein binnen de geneeskunde waarvoor de applicatie kan gebruikt worden (bv. radiografie, dermatologie).
Specifieke doel	De symptomen, aandoeningen, diagnoses, onderzoeken waarvoor de applicatie kan gebruikt worden (bv. psoriasis, hoofdpijn, spirometrie).
Samenvatting methodologie	Een samenvatting van de gebruikte methodologie.
Samenvatting resultaten	Een samenvatting van de resultaten.

Stap 8: interpretatie en rapportage van de resultaten

Naast een overzichtelijke weergave van de bovengenoemde data-extractie, werd een subjectieve inschatting van de kwaliteit van de methodologie/resultaten genoteerd. Dit wordt besproken in de discussiesectie.

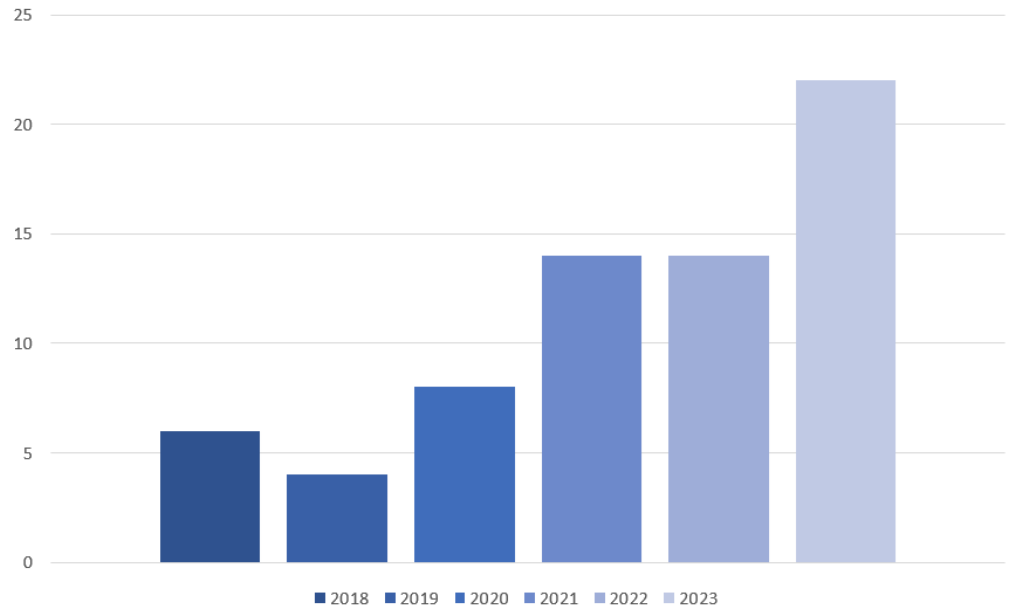
Resultaten

Overzicht van resultaten

Een overzicht van de geëxtraheerde data uit de verschillende studies wordt weergegeven in Tabel 5. In de eerste kolom wordt het doel van de studie, het type AI, de functie van de AI-toepassing, het subdomein waarbinnen de AI-toepassing gebruikt kan worden en het specifieke doel van de studies toegelicht (zie ook Tabel 4 voor definities). In de tweede kolom wordt een korte samenvatting van de gebruikte methodologie weergegeven. Tot slot worden de hoofdbevindingen toegelicht. Een grote variabiliteit werd opgemerkt in de manier waarop methodologie en resultaten gerapporteerd werden in de geïncludeerde studies.

Publicatiejaar van de studies

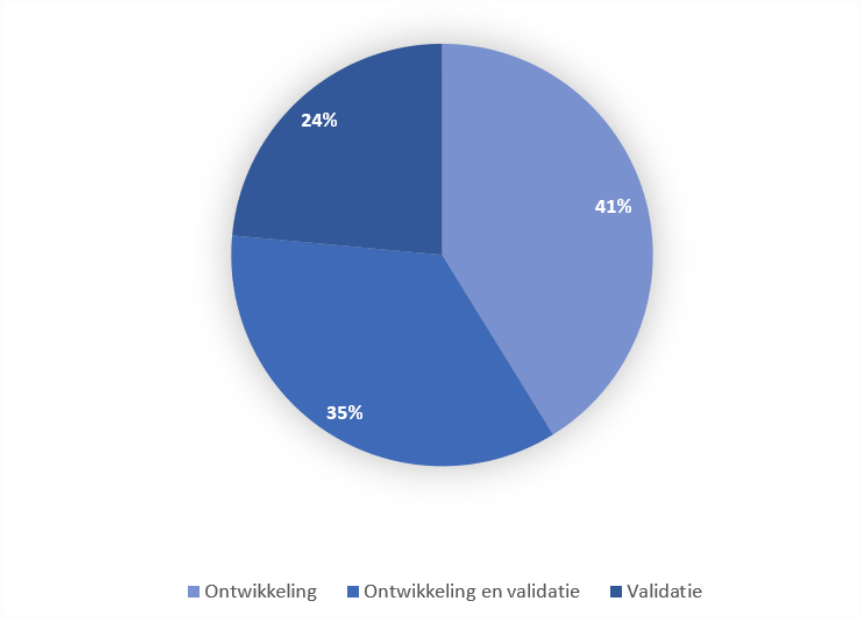
In Figuur 3 wordt gevisualiseerd in welk jaar de geïncludeerde studies gepubliceerd werden.



Figuur 3. Een overzicht van het aantal studies per jaar.

Doel van de studies

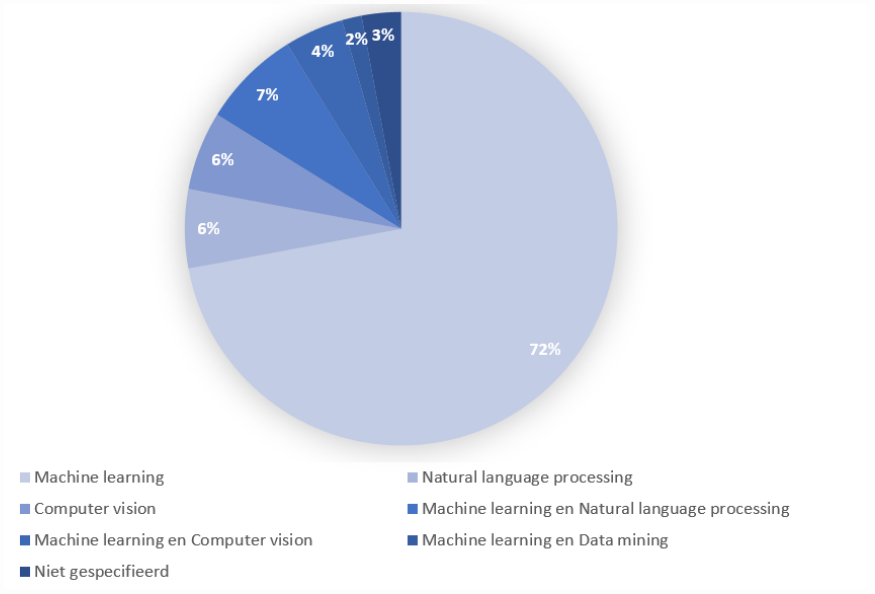
In Figuur 4 wordt een overzicht van de doelen van de geïncludeerde studies weergegeven. 41% van de studies had als doel om AI-toepassingen te ontwikkelen, 24% van de studies om de AI-toepassingen te valideren, en 35% van de studies beoogde een combinatie van beide. Binnen het validatieonderzoek, waren er slechts tien prospectieve studies (15%) die een AI-toepassing onderzochten (41–50). Zes van deze studies werden gepubliceerd in 2022-2023.



Figuur 4. Een overzicht van de doelen van de studies.

Type artificiële intelligentie

In Figuur 5 worden de verschillende AI-types gebruikt in de geïncludeerde studies weergegeven. De meerderheid van de AI-toepassingen maakte gebruik van *machine learning* (72%), *natural language processing* (6%) of *computer vision* (6%).

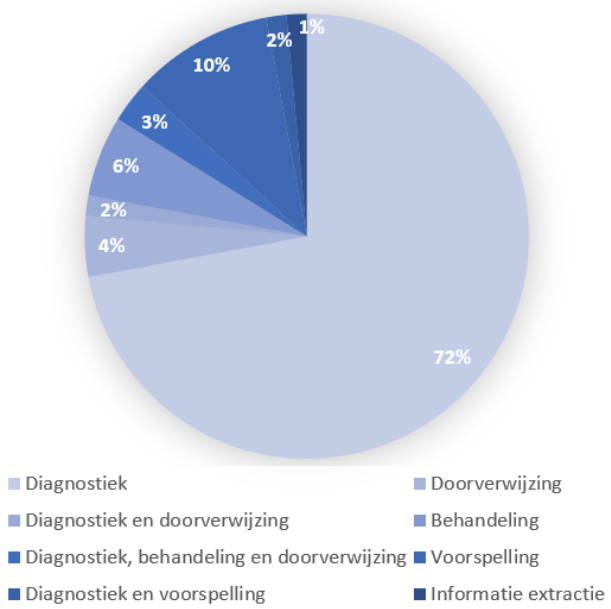


Figuur 5. Een overzicht van de AI-types.

Functie artificiële intelligentie

In Figuur 6 worden de verschillende AI-functies die onderzocht werden in de geïncludeerde studies gevisualiseerd. De meerderheid van de AI-toepassingen (72%) beoogden om diagnostiek te ondersteunen. 10% van de AI-toepassingen hadden tot doel om voorspellingen te doen over zaken in de toekomst, onder meer het voorspellen van hypertensie (51,52), hypoglycemie bij type 2 diabetes (53), suïcide (54,55), vallen bij geriatrische patiënten (56), COVID-19 gerelateerde hospitalisaties (57) en therapietrouw en kwaliteit van shared decision making bij astmapatiënten (58).

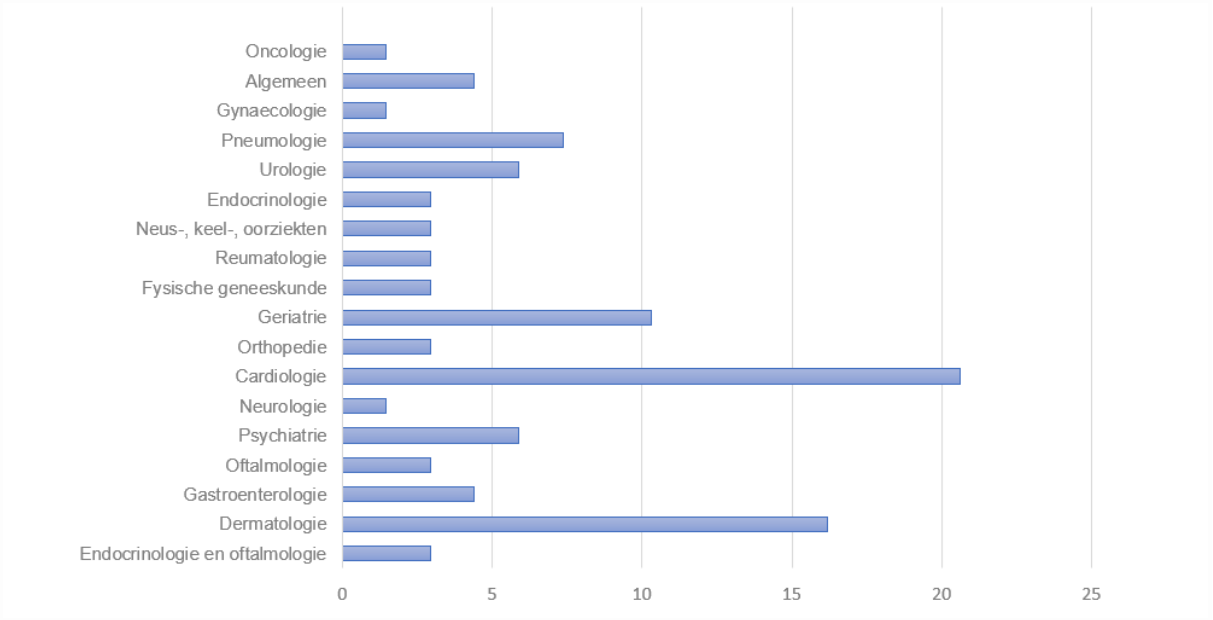
De grote meerderheid van de AI-toepassingen handelden over specifieke symptomen of aandoeningen. Meer bepaald waren er AI-toepassingen uitgewerkt rond hartfalen en/of hartritmestoornissen (47,49,59–62), voorkamerfibrillatie (46,63–65), hypercholesterolemie (66,67), huidaandoeningen en huidletsels (43,44,68–71), melanoma (72,73), psoriasis capitis en seborroïsche keratose (74), vitiligo (75), dementie (76–80), milde cognitieve achterstand (73), kanker (82), maag- en slokdarmkanker (83), pancreascarcinoom (84), colorectale kanker (85), endometriumcarcinoom (86), longkanker (87), cystitis (88), urineweginfectie (45,89), retinopathie (90,91), ptose (92,93), lage rugpijn (50,94), clusterhoofdpijn, migraine en spanningshoofdpijn (95), fracturen (96,97), spondylitis ankylosans (98), astma en/of COPD (58,99), COVID-19 (57), depressie (100) en posttraumatische stresssyndroom (101).



Figuur 6. Een overzicht van de AI-functies.

Subdomein geneeskunde

In Figuur 7 wordt het percentage van de geïncludeerde studies binnen een bepaald subdomein weergegeven. De meerderheid van de AI-modellen kenden hun toepassing binnen cardiologie (21%), dermatologie (16%) en geriatrie (10%).



Figuur 7. Percentage van de geïncludeerde studies (x-as) binnen een bepaald subdomein van de geneeskunde (y-as).

Tabel 5. Overzicht resultaten

Studie **Doel studie** **Type AI** **Functie AI** **Subdomein** **Specifiek doel**	**Methodologie**	**Resultaten**
Abascal Azanza et al. (92) O en V CV en ML Diagnostiek Ptose Oftalmologie	De dataset bestond uit 250.534 afbeeldingen van 10.555 patiënten uit een derdelijns oftalmologische afdeling, verzameld tussen 1993 en 2018. Alle afbeeldingen werden beoordeeld door één expert. Uit deze dataset werden 2000 afbeeldingen gekozen van 1000 patiënten met ptose en 1000 controles zonder ptose. Hierop werd de AI-toepassing getraind. Een tweede dataset werd verzameld uit openbare databanken, zoals Google Images en Realself. Deze dataset bestond uit 567 patiënten met 1389 afbeeldingen. Na beoordeling op geschiktheid en exclusie werd een steekproef gegenereerd met 11 afbeeldingen van ptose en 89 afbeelding zonder ptose. Op de eerste dataset werd de toepassing getraind en geëvalueerd. Met de steekproef uit de tweede dataset werden de prestaties van de toepassing vergeleken met 9 artsen, waaronder 3 huisartsen.	Evaluatie op de eerste dataset toonde aan dat het AI model een AUC van 0,918 haalde voor het maken van een onderscheid tussen patiënten met en zonder ptose. Het model vertoonde een sensitiviteit van 98% en een specificiteit van 88%. In de vergelijkende studie met 9 artsen was de AUC voor deze tweede dataset 0,700. Het model vertoonde een sensitiviteit van 54% en een specificiteit van 85%. Na vergelijking van deze prestaties met de hoogste en laagste AUC's van de artsen, bleken de verschillen (respectievelijk 0,142 en 0,0618) statistisch niet-significant te zijn (respectievelijk p = 0,6328 en p = 0,2220).
Akyea et al. (66) O ML Diagnostiek Familiale hypercholesterolemie Cardiologie	De dataset was deel van de Clinical Practice Research Datalink (CPRD), een verzameling van medische gegevens uit 836 huisartspraktijken met 5,2 miljoen bruikbare patiëntendossiers. De dataset bestond uit 4.027.775 personen met minstens 1 cholesterolmeting, waarvan 7928 (0,2%) een diagnose van familiale hypercholesterolemie hadden. 75% (n = 3.020.832) werd willekeurig geselecteerd als trainingsset, en de overige 25% (n = 1.006.943) toegewezen als validatieset. De gegevens bestonden uit 45 variabelen, zoals triglyceriden, totaal cholesterol, LDL, leeftijd, (familiale) voorgeschiedenis, etc.	Vier van de vijf Machine Learning modellen hadden een gelijkaardige hoge voorspellende nauwkeurigheid (AUC > 0.89), één model had een AUC van 0.81. Het aantal patiënten waar het machine learning model met een hoge probabiliteit de diagnose voorspelde varieerde van 0,73% voor het ensemble-model gebruikt tot 10,16% voor het deep learning model. Deze casussen vereisen verdere klinische evaluatie. De vijf machine learning modellen toonden verschillende resultaten op variabelen zoals sensitiviteit, specificiteit en

	Er werden 5 verschillende vormen van *Machine Learning* getest op het detecteren van familiale hypercholesterolemie aan de hand van 45 variabelen.	positieve en negatieve predictieve waarde. Naargelang de voorkeur kan men opteren voor een bepaald model. Zo heeft het ensemble-model een relatief lage sensitiviteit (30.5%) en hogere positief voorspellende waarde (15.5%), terwijl het deep learning voor die metrieken respectievelijk 72.6% en 2,8% haalt. De keuze voor de eerste houdt een efficiëntiewinst in, met mogelijk het missen van enkele gevallen.
Anderson et al. (96) V ML Diagnostiek Fracturen Orthopedie	De gebruikte dataset bestond uit casussen afkomstig uit 12 verschillende ziekenhuizen en praktijken. 175 Casussen werden weerhouden. De casussen bestonden uit röntgenfoto's (RX) van 12 anatomische gebieden (enkel, sleutelbeen, elleboog, dijbeen, onderarm, heup, opperarmbeen, knie, bekken, schouder, scheenbeen en kuitbeen, en pols). Een patiënt kon meerdere RX's hebben van verschillende gebieden. De casussen werden tweemaal beoordeeld door de 24 deelnemende artsen, van wie 4 huisartsen. De tijd tussen de 2 sessies was minimaal 28 dagen. Bij de eerste sessie werden de helft van de casussen beoordeeld door de artsen met hulp van de toepassing, bij de tweede helft zonder. In de tweede sessie werd dit omgekeerd gedaan. De gouden standaard voor de radiografieën werd bepaald door de onafhankelijke evaluatie van 3 orthopedische chirurgen/radiologen.	De toepassing verbeterde de nauwkeurigheid van de artsen bij het diagnosticeren van musculoskeletale fracturen (zonder AI AUC: 0,903 (BI 0,890 tot 0,916); met AI AUC: 0,944 (BI 0,934 tot 0,954)). De sensitiviteit van de artsen steeg met AI tegenover zonder AI (AI: 90% (BI 88% tot 92%); Zonder AI: 82% (BI 79% tot 84%)), en idem voor de specificiteit (AI: 92% (BI 91% tot 93%); zonder AI: 89% (BI 88% tot 90%)). Artsen met beperkte training in het interpreteren van RX (huisartsen, arts-assistenten en spoedartsen) vertoonden de grootste verbeteringen in sensitiviteit en specificiteit wanneer ze werden ondersteund door toepassing. Clinici met beperkte training in deze beeldvorming zagen een hoger percentage fracturen over het hoofd indien ongeholpen door AI in vergelijking met radiologen (foutenpercentage artsen met beperkte beeldvorming training: 20% (BI 17% tot 24%); foutenpercentage radiologen: 14% (BI 9% tot 19%)). Echter, wanneer ondersteund door de toepassing, daalde het foutenpercentage van clinici met beperkte training, resulterend in een vergelijkbaar foutenpercentage als radiologen (foutenpercentage voor clinici met beperkte training: 9% (BI 7% tot 12%); foutenpercentage voor radiologen: 10% (BI 6% tot 15%)).
Bennis et al. (59)	Data van 8543 patiënten van 70 jaar en ouder (1 tot 2 jaar voor de diagnose van hartfalen) en 8543 controle casussen werd retrospectief verkregen uit de	Wanneer enkel de demografische gegevens beschikbaar waren, werd een AUC van 0,692, (BI 0,677 tot 0,706)

O ML Diagnostiek Hartfalen Cardiologie	*Nivel Primary Care Database*. De gegevens werden verkregen in de periode van 1 januari 2012 tot 31 december 2019. Er wordt gebruik gemaakt van twee soorten data: demografische gegevens en codes die informatie bevatten rond chronische ziekten, klachten, diagnostiek en medicatie. De patiënten en deze gegevens werden verdeeld in een training- en testset. Datasets met deze demografische gegevens, de aanwezigheid van codes (niet-sequentieel, geen informatie rond tijd) en opeenvolgende codes (sequentieel, met temporele informatie) werden gecreëerd. Logistische regressie, *random forest* en *XGBoost*-modellen werden getraind. Het voorspelde resultaat was de aanwezigheid van hartfalen (HF) na 1 jaar. De verhouding geval:controlegroep in de testset kwam overeen met de werkelijke incidentiecijfers (1:45).	behaald. Het toevoegen van niet-sequentiële informatie in combinatie met een logistiek regressiemodel presteerde het beste en verbeterde de prestaties aanzienlijk (AUC 0,772(BI 0,759 tot 0,785), p<0,001). Verdere toevoeging van sequentiële informatie, wat temporele informatie toevoegt, veranderde de prestaties niet significant (AUC 0,767, BI 0,754 tot 0,780, p=0,07). Met het beste model kon men concluderen dat er 5.99 vals positieve resultaten waren ten opzichte van 1 echt positief resultaat.
Briggs et al. (83) O ML Diagnostiek Maag- en slokdarmkanker Gastroenterologie	De gebruikte dataset in deze studie was afkomstig van de *General Practice Research Database* (GPRD) in het Verenigd Koninkrijk en werd verkregen uit een case-control studie uit 2013 over risicofactoren voor maag- en slokdarmkanker. Deze gegevens, inclusief informatie van ongeveer 4,8 miljoen patiënten, werden gebruikt om het *ogRAT*-model te ontwikkelen. De steekproef bevatte gegevens van 40.348 patiënten (7471 gevallen en 32.877 controles), geselecteerd op basis van tumor diagnosecodes. Inclusiecriteria vereisten een diagnose na het jaar 2000, een leeftijd van ≥40 jaar en beschikbare consultaties in het jaar voorafgaand aan de diagnose. De verzamelde medische informatie omvatte demografische gegevens, ICPC-gecodeerde symptomen en laboresultaten. De gemiddelde leeftijd van de patiënten was 73 jaar, met diagnoses van slokdarm- en maagkanker in respectievelijk 65,3% en 34,7% van de gevallen. Vijf machine learning modellen werden gekozen en uitgewerkt. Het probleem werd geformuleerd als een binaire classificatietaak tussen 'kanker' en 'geen kanker', met bijbehorende kans op kanker.	Drie van de vijf modellen (logistische regressie, *support vector machine* en *extreme gradient boosted decision tree*) behaalden de beste prestaties (accuraatheid 0.89, AUROC 0.87). Deze modellen presteerden beter dan de huidige Britse risico assessment tool *(ogRAT)*. Ook Machine learning identificeerde meer kankerpatiënten dan *ogRAT*: 11% meer, met weinig tot geen effect op vals positieven of 25% meer met een lichte stijging in vals positieve casussen.
Calcagno et al. (41) V	De steekproef bestond uit mannen en vrouwen met minstens drie risicofactoren voor hartaandoeningen, die een huisartsenpraktijk bezochten voor algemene gezondheidsredenen, zoals jaarlijks lichamelijk onderzoek, in een typische huisartsenpraktijk. De beoogde steekproefgrootte was minimaal 500 patiënten,	De AI-toepassing *CHART* verminderde de vals negatieve percentages met 15.8% en de vals positieve percentages met 5% in vergelijking met de standaard ECG-apparaten. Huisartsen die *CHART* gebruikten, zagen het percentage

Niet gespecifieerd Doorverwijzing ECG Cardiologie	waarvan ten minste 15% als gezond (zonder risicofactoren) werd beschouwd. De leeftijdsverdeling was als volgt: 61 patiënten tussen 20 en 40 jaar, 130 tussen 41 en 55 jaar, 156 tussen 56 en 65 jaar, 148 tussen 66 en 75 jaar en 52 boven de 76 jaar. De gewichtsverdeling was: 6,2% ondergewicht, 35,8% normaal gewicht, 34,7% overgewicht en 23,3% obees. Er zouden 550 patiënten geïncludeerd zijn. 550 patiënten werden gezien door 2 artsen, 1 arts had toegang tot een standaard ECG toestel met automatische analyse, de andere arts had toegang tot het *Cardio-HART(CHART)* toestel. De uitkomst van de studie was het al dan niet verwijzen van een patiënt naar de cardioloog voor verder onderzoek. De gegevens werden ook geïnterpreteerd door een cardioloog, wat de gouden standaard vormt.	positieve diagnoses aanzienlijk toenemen, van ~10% naar ~26%, en het onzekerheidspercentage aanzienlijk afnemen, van ~31% naar ~1,9%.
Cha et al. (102) O en V ML Diagnostiek Otoscopie Neus-, keel-, oorziekten (NKO)	De dataset bestond uit 22,707 otoscopische beelden van patiënten uit het *Severance Hospital* in Seoul, Zuid-Korea. Na exclusie van ongeschikte foto's werden 7500 beelden geselecteerd. Deze werden in 6 categorieën ingedeeld op basis van aandoeningen aan het trommelvlies en de gehoorgang. 6900 beelden werden gebruikt voor de training van de modellen. Van de overige 600 beelden werden twee steekproeven van 300 beelden gegenereerd voor het testen van de toepassing. Een eerste dataset, de gebalanceerde dataset, waarbij van elke diagnose een gelijk aantal beelden werden gegeven en een tweede dataset waarbij de beelden afhankelijk waren van de prevalentie van de diagnose in de volledige dataset, de ongebalanceerde dataset. De beelden van de twee bovenstaande steekproeven werden voorgelegd aan 14 artsen: 6 NKO-artsen (2 NKO-artsen, 4 NKO-artsen in opleiding), 8 niet-NKO-artsen met ervaring in otoscopie (2 spoedartsen, 6 huisartsen). Alle antwoorden van de artsen en de ML-modellen werden naast elkaar gelegd om de nauwkeurigheid te vergelijken.	De modellen hadden consistent hoge nauwkeurigheid (gebalanceerde testset: gemiddeld 77,14%, SD 1,83%; ongebalanceerde testset: gemiddeld 82,03%, SD 3,06%), gelijkaardig aan die van NKO-artsen (gebalanceerd: gemiddeld 71,17%, SD 3,37%; ongebalanceerd : gemiddeld 72,84%, SD 6,41%) en veel beter dan die van niet-NKO-artsen (gebalanceerd: gemiddeld 45,63%, SD 7,89%; ongebalanceerd: gemiddeld 44,08%, SD 15,83%). De modellen hadden echter last van problemen met de mate waarin de data gebalanceerd was (gebalanceerde testset: gemiddeld 77,14%, SD 1,83%; ongebalanceerde testset: gemiddeld 82,03%, SD 3,06%). Dit werd verbeterd door gegevens toe te voegen, vooral voor diagnoses met een lage incidentie, maar de klassen van zeldzame ziekten hadden nog steeds een lage nauwkeurigheid per klasse. Binnen de artsen werd een hoge variabiliteit vastgesteld, ondanks dat ze minder beïnvloed werden door de prevalentie,(ML-modellen: kappa=0,83, SD 0,02; NKO-artsen: kappa=0,60, SD 0,07). ML-modellen hebben dus een consistente en hoge nauwkeurigheid, terwijl ze enkel rekening houden met het gegeven beeld en een voorkeur voor prevalentie vertonen,

		terwijl artsen wisselende prestaties leveren, maar geen voorkeur voor prevalentie vertonen en ook extra informatie kunnen overwegen die geen afbeeldingen zijn.
Chen et al. (51) O ML Voorspelling Hypertensie Cardiologie	De dataset werd gehaald uit de elektronische patiëntendossiers van 47 gezondheidscentra in het Pudong-district van Shanghai. Patiëntendossiers, onderzoeken en andere gerelateerde gegevens van patiënten ouder dan 35 jaar (tussen 2017 tot 2019) werden geïncludeerd in de dataset. In totaal ging het om 40.261 patiënten. 20 Kenmerken werden geselecteerd om het model te trainen uit de 127 variabelen: leeftijd, diabetes, proteïnurie, BMI, zelfbeoordeling van de gezondheid bij ouderen, creatinine, systolische en diastolische bloeddruk, buikomtrek, rookgedrag, totaal cholesterol, LDL-C, HDL-C, triglyceriden Alcoholgebruik, glucosewaarden, ureum, frequentie van fysieke activiteit, tijd besteed aan fysieke activiteit, hoog zoutgebruik. Er werden 3 steekproeven gebruikt, een trainingset, validatieset en testset. Er werden drie modellen onderling vergeleken.	Het *XGBoost* model slaagde er beter in dan de andere twee modellen (r*andom forest*, logistische regressie) om het risico op hypertensie te voorspellen. Dit weerspiegelde zich in de behaalde sensitiviteit (0.864), specificiteit (0.488), positief voorspellende waarde (0.735), negatief voorspellende waarde (0.686), accuraatheid (0.722) en AUC (0.765) in de testset.
Chung et al.(97) O en V ML Diagnostiek Proximale humerusfractuur Orthopedie	De dataset bestond uit 1891 afbeeldingen van normale schouders (n=515) en 4 types proximale heupfracturen (n=1376), geclassificeerd door drie specialisten. De afbeeldingen werden verzameld in zeven verschillende ziekenhuizen uit Korea. Een AI-model werd getraind om breuken te diagnosticeren op 9/10 van de dataset. De overige 1/10 van de dataset werd beschouwd als testset. Het AI-model werd vergeleken met de resultaten van 28 huisartsen, 11 algemene orthopedisten en 19 orthopedisten gespecialiseerd in de schouder.	Het AI-model had een betere prestatie dan huisartsen en algemene orthopedisten, en een gelijkaardige prestatie aan orthopedisten gespecialiseerd in de schouder. Het AI-model was vooral beter op het vlak van het vaststellen van complexe breuken. Het AI-model had een sensitiviteit van 0.99 (BI 0.99–1.00), en een specificiteit van 0.97 (BI 0.97–0.98), in vergelijking met een sensitiviteit van 0.82 (BI 0.78-0.87) en specificiteit van 0.94 (0.93-0.96) bij huisartsen.
Crutzen et al. (53) O ML Voorspelling	De AI-modellen werden ontwikkeld op basis van de '*Groningen Initiative to ANalyse Type 2 diabetes Treatment (GIANTT)*' medische databank om hyperglycemie (=eerste hypoglycemische aanval tussen 2007-2013) te voorspellen. Demografische variabelen en medische gegevens werden gebruikt om de AI-modellen te trainen. De steekproef bestond uit 13876 patiënten	Het meest accurate model had een AUC van 0.71 op basis van 10 geïncludeerde variabelen (waaronder geslacht, leeftijd, medicatiegebruik, insulinegebruik etc.). Het toevoegen van meer variabelen had geen meerwaarde op de resultaten van de modellen.

Hypoglycemie bij type 2 diabetespatiënten Endocrinologie	gediagnosticeerd met type 2 Diabetes door 180 verschillende huisartsen in Nederland. 80% van de data werd gebruikt als trainingset. De overige 20% werd gebruikt als testset.	
Cruz et al. (42) O en V NLP en ML Diagnostiek, behandeling en doorverwijzing Richtlijnen over 18 klachten (hoofdpijn, angst, lage rugpijn etc.) Algemene geneeskunde	Het doel van deze studie was de validatie van een AI-toepassing die aanbevelingen gaf aan artsen tijdens consultatie. Deze aanbevelingen waren gebaseerd op de richtlijnen. Eerst werden de richtlijnen vertaald in regels. Met deze regels werden consultaties van artsen uit spoedafdelingen en huisartsgeneeskunde geëvalueerd op het volgen van de richtlijnen. Uit deze resultaten werden 18 richtlijnen gevonden waarbij minder dan 90% naleving was. Tot slot werd de toepassing geïmplementeerd en geëvalueerd hoe de toepassing de naleving van de richtlijnen beïnvloedde. De gegevens kwamen uit 24 praktijken met een totaal van 86 arsten (huisartens en pediaters).Voor de initiële evaluatie van de naleving werden 2.5 miljoen klinische documenten gebruikt, vekregen tussen 01/01/2016 en 31/05/2017. Voor de implementatie van de AI toepassing werden er 345,000 documenten gebruikt van 27/11/2017 tot 31/12/2017.	Het naleven van de richtlijnen verbeterde bij 8 van de 18 richtlijnen na implementatie van de tool. Er werd echter slechts een significante verbetering gevonden bij 3 richtlijnen. De andere 10 richtlijnen konden niet geëvalueerd worden, door te weinig data. Verder keken de auteurs naar het aantal meldingen. Zo werd er tijdens de implementatieperiode, een totaal van 534 aanbevelingen gegeven, elk van de 86 artsen kreeg dus gemiddeld 1.8 meldingen per dag. Tot slot evalueerde men de ervaring van de artsen. De algemene indruk was zeer positief en men vond de toepassing een meerwaarde. Ook stoorden de meldingen de artsen niet tijdens het consulteren.
DelPozo-Banos et al. (54) O ML Voorspelling Suïcide Psychiatrie	De *Secure Anonymised Information Linkage Databank UK* werd gebruikt in deze studie. Data verzameld bij huisartsen en ziekenhuizen van personen die zelfmoord pleegden tussen 2001 en 2015 (n=2604), en controle patiënten (n=52080) werden geïncludeerd. Verschillende variabelen werden opgenomen in het model (bv. aantal consultaties, psychiatrische diagnoses etc.). Het AI-model werd getraind om te differentiëren tussen de twee groepen van patiënten. 8/10 van de data werd gebruikt als training set, 1/10 als testset en 1/10 als validatieset.	De verschillende versies van het model werden geëvalueerd op hun mogelijkheid casussen met hoog risico op suïcide te identificeren. Het beste model behaalde een sensitiviteit van 0.646, een specificiteit van 0.819 en een AUC van 0.80.
Dhanda et al. (88)	De dataset(n=472) van huisartsen (8 huisartsen) van de afdeling huisartsengeneeskunde van '*University of Kansas Medical Center*'. Patiënten	Het model voorspelde dat 119 culturen pathogeen zouden zijn en 134 niet-pathogeen. Voor de voorspelde pathogene

V ML Behandeling Cystitis Urologie	werden geëxcludeerd indien er symptomen waren van een ernstige ziekte (pyelonefritis/sepsis.). Er werden 253 patiënten geïncludeerd. De variabelen die werden gebruikt voor de ML-modellen waren leeftijd, geslacht, dipstickanalyse, symptomen zoals dysurie en buikpijn, en de geschiedenis van een urineweginfectie. Verschillende soorten ML-modellen werden onderzocht op de mogelijkheid om het resultaat van urineculturen te voorspellen op basis van de variabelen. Het doel van deze toepassing is om de indicatie tot antibioticagebruik te verbeteren.	culturen raadde de beslissingsregel geen verandering aan in de besluitvorming over antibiotica. In de voorspelde niet-pathogene culturen waren er 15 gevallen waarin het model voorstelde om antibiotica niet voor te schrijven in situaties waarin artsen antibiotica wel voorschreven. 14 van deze 15 gevallen hadden een negatieve urinekweek, terwijl slechts 1 een positieve cultuur had. Soortgelijke resultaten werden verkregen met behulp van de andere modellen. De voorstellen van het model zouden de incidentie waarvoor geen antibiotica werd voorgeschreven, bij patiënten met negatieve urinekweken, verhogen van 81,9% naar 89,4%, terwijl de incidentie waarvoor antibiotica (ten onrechte) niet werd voorgeschreven aan patiënten met positieve urinekweken slechts zou toenemen van 23,1% naar 24,6%.
Dormosh et al. (56) O NLP en ML Voorspelling Vallen Geriatrie	De databank '*The Academic General Practitioner's Network at the Academic Medical Center*' werd gebruikt in deze studie. Deze databank bestaat uit data verzameld bij 50 huisartsen in Nederland tussen 2012 en 2019 van 36470 patiënten (>65 jaar). Verzamelde gegevens bestonden uit demografische gegevens, diagnosen, medicatie en ongestructureerde notities. Deze gegevens werden gebruikt om patiënten die vallen te identificeren (n=4778) aan de hand van natural language processing en ML-modellen.	Het beste model combineerde gestructureerde gegevens met ongestructureerde notities en had een AUC van 0.718 (BI 0.708–0.727).
Dros et al. (103) O ML Diagnostiek Sjörgen syndroom Reumatologie	De steekproef bestond uit 1411 patiënten met Sjörgen syndroom en 929179 patiënten zonder uit Nederland. De dataset bestond uit data uit eerstelijnszorg en data uit gespecialiseerde zorg tussen 2012 en 2017. Twee verschillende ML-modellen werden ontwikkeld om patiënten om patiënten met Sjögren syndroom te identificeren op basis van leeftijd, geslacht, voorgeschiedenis, symptomen, voorschriften en aantal consultaties. 75% van de data werd gebruikt als trainingset 25% van de data werd gebruikt als testset.	De twee modellen hadden een AUC van 0.82 en 0.84, een sensitiviteit van 0.723 en 0.701, een specificiteit van 0.74 en 0.779. De negatief voorspellende waarde was 99.9% voor beide modellen. De positief voorspellende waarde was echter maar 0.4% en 0.5%.
Ellertsson et al. (95)	De dataset bestond oorspronkelijk uit 15024 klinische notities verzameld van 13682 consultaties voor hoofdpijn (clusterhoofdpijn, migraine en	Het AI-model deed het beter dan beoordeling door huisartsen op vlak van sensitiviteit, positief voorspellende waarde, en

O en V ML Diagnostiek Clusterhoofdpijn, migraine en spanningshoofdpijn Neurologie	spanningshoofdpijn). Dit verzameld in 15 verschillende huisartspraktijken in Ijsland tussen 2006 en 2020. Na inclusie/exclusie werden er 800 klinische notities willekeurig geselecteerd. Deze notities werden gelabeld met binaire variabelen (aan- of afwezigheid van een symptoom). 75% Van de data behoorde tot de trainingset, 12.5% tot de validatieset en 12.5% tot de testset. De resultaten van het AI-model op de testset werd vergeleken met de resultaten van zes huisartsen, waarvan 3 in opleiding.	AUROC. Het model toont een sensitiviteit van 0.95, specificiteit van 0.88 en positief voorspellende waarde van 0.94. Dit in vergelijking met de artsen, respectievelijk 0.86(0.78-0.9), 0.9(0.78-0.95) en 0.86(0.76-0.9).
Escalé-Besa et al. (43) V CV Diagnostiek Huidlaesies Dermatologie	Deze prospectieve studie onderzocht 100 patiënten met huidletsels die de huisarts consulteerden in Catalonië tussen juni en oktober 2021. De huidproblemen werden eerst onderzocht door de huisarts(n=11). Zij gaven een top 3 van diagnoses. Vervolgens werd een foto genomen en werd een AI-tool gebruikt. De huisartsen evalueerden dan de meerwaarde van het resultaat van de toepassing. De toepassing gaf een lijst met vijf mogelijke diagnoses, in volgorde van waarschijnlijkheid. Dezelfde foto werd naar een 2 dermatologen gestuurd. Indien er geen overeenkomst was tussen de dermatologen, werd een derde arts ingeschakeld of het resultaat van histologie beschouwd als gouden standaard.	De accuraatheid van het ML-model in Top-1 was 0,39 (0,29–0,49) in vergelijking met 0,72 (0,62–0,80) voor dermatologen en 0,64 (0,54–0,73) voor huisartsen. Deze waarden nemen toe wanneer Top-3 wordt beoordeeld, met een accuraatheid van 0,61 (0,51–0,71) voor het ML-model en bereikt 0,72 (0,62–0,80) voor Top-5. Alle waarden van accuraatheid van het ML-model waren lager dan die van de artsen. Er waren echter 18 gevallen waarin het model de ziekte niet kon herkennen, omdat het niet was getraind op de specifieke diagnose. Daarom werd een subanalyse uitgevoerd waarin alleen de 82 gevallen werden opgenomen die overeenkwamen met de diagnoses waarmee het model was getraind, waarna de accuraatheid toenam tot 0,48 (0,37–0,59) in Top-1, tot 0,75 (0,66–0,85) in Top-3 en tot 0,89 (0,79–0,95) in Top-5.
Ford et al. (78) O ML Diagnostiek Dementie Geriatrie	De '*Clinical Practice Research databank'* werd gebruikt, dit is een databank van gegevens uit de eerstelijnszorg in Engeland. Patiënten met (n=45,650) en gekoppelde controles zonder dementie(n=45,650) werden geïncludeerd (>65 jaar). De verzamelde data bestond uit verschillende variabelen verzameld vijf jaar voor de diagnose, zoals medische diagnoses, doorverwijzingen naar specialisten, informatie over de levensstijl, vallen, bloeddruk etc. Vijf	Vier van de vijf modellen behaalden een AUC van 0.75. Het logistische regressiemodel had een specificiteit van 0.752, een sensitiviteit van 0.602 en een positief voorspellende waarde van 0.156.

	AI-modellen werden ontwikkeld op basis van de trainingset (80% van dataset), en de AUC werd bepaald in de testset (20% van de dataset).	
Ford et al. (77) O ML Diagnostiek Dementie Geriatrie	De '*Clinical Practice Research databank'* werd gebruikt, dit is een databank van gegevens uit de eerstelijnszorg in Engeland. Patiënten met (n= 47,601) en zonder (n= 47,601) dementie werden geïncludeerd (>65 jaar) tussen 2000 en 2012. Er werden ook subgroepen geïdentificeerd, namelijk patiënten met Alzheimer (n=13,452) en vasculaire dementie (n=10,870) De verzamelde data bestond uit verschillende variabelen, zoals bijvoorbeeld medische diagnoses, doorverwijzingen naar specialisten, informatie over de levensstijl etc. 67% van de dataset werd gebruikt als trainingset, 33% als testset. Er werden verschillende modellen ontwikkeld en getest, op data vijf, vier, drie, twee en één jaar voor de diagnose. Daarnaast werden ook modellen ontwikkeld om dementie in het algemeen te voorspellen, en modellen om subtypes te voorspellen.	Modellen die enkel gebruik maakten van data vijf tot twee jaar voor de diagnose toonden zwakke resultaten (AUC 0.55-0.65), maar werden beter naargelang de duur tussen meting en diagnose korter werd. Het model dat data gebruikte van één jaar voor de diagnose had een AUC van 0.84. Gelijkaardige resultaten werden gevonden bij de voorspelling van de subtypes. In het laatste jaar voor de diagnose was de onderscheiding van het Alzheimer-model zeer goed (AUC 0,89), en het Vasculaire model ook (AUC 0,85). Op de andere jaren waren de resultaten echter zwak voor beide modellen (AUC 0,54–0,67).
Ford et al. (76) O ML Diagnostiek Dementie Geriatrie	De '*Clinical Practice Research databank*' werd gebruikt, dit is een databank van gegevens uit de eerstelijnszorg in Engeland. Patiënten met (n= 46713) en zonder (n= 46713) dementie werden geïncludeerd (>65 jaar) tussen 2000 en 2012. De verzamelde data bestond uit verschillende variabelen, zoals medische diagnoses, doorverwijzingen naar specialisten, informatie over de levensstijl etc. Daarnaast werden ook klinische notities van de huisartsen verzameld, gebundeld in negen concepten (bv. vergeetachtig zijn, verward zijn, familiale zorgen etc.). Drie AI-modellen werden ontwikkeld op basis van de trainingset, en de AUC werd bepaald in de testset met als doel patiënten met dementie te identificeren.	De drie modellen hadden een AUC variërend tussen 0.87 en 0.90. Wanneer de negen concepten werden toegevoegd aan de modellen, was de AUC 0.90-0.94. Ook steeg de sensitiviteit met ongeveer 0.08 na toevoeging van deze modellen, 0.76-0.78 naar 0.8-0.85. Dit zonder verandering van de specificiteit.
Giavina-Bianchi et al. (73) O CV	Er werden twee modellen ontwikkeld: één voor foto's gemaakt met de smartphone en één voor foto's gemaakt door dermatoscopisch onderzoek. De dataset voor de ontwikkeling van het dermatoscopisch model bestond uit 26,342 afbeeldingen van de *ISIC2019* en *PH* databanken. De dataset bestond ook uit 4000 afbeeldingen van melanomen gecreëerd door AI. Twee dermatologen beoordeelden alle afbeeldingen.	De accuraatheid van het dermatoscopisch model voor het detecteren van melanomen was 89.3% met een sensitiviteit van 91% en een specificiteit van 89%. De accuraatheid van het klinische model voor het detecteren van melanomen was 84.7% met een sensitiviteit van 0.89 en een specificiteit van 0.83. Beide modellen hadden aan AUC hoger dan 0.90.

Diagnostiek, doorverwijzing en behandeling Melanoma Dermatologie	De dataset voor de ontwikkeling van het klinisch model (met gebruik van smartphone foto's) maakte deel uit van het *HIAE's teledermatology project*. De steekproef bestond uit 14,000 afbeeldingen genomen met een smartphone door huisartsen. Een foto van het huidletsel werd verwerkt door een smartphone app. De App werd gebruikt door huisartsen en gaf vervolgens verschillende zaken aan: de voorspelde waarschijnlijkheid dat de huidleasie kwaadaardig was, de waarschijnlijke diagnose gebaseerd op die waarschijnlijkheid, en een suggestie voor verdere behandeling, en een heatmap. Uiteindelijk werden 3083 afbeeldingen geselecteerd voor de steekproef. Twee dermatologen beoordeelden ook alle afbeeldingen. Voor beide modellen gebruikte men 80% van de dataset als de trainingset, 10% als de validatieset en 10% als de testset.	
Gregoor et al. (44) V ML Diagnostiek Maligne huidletsels Dermatologie	Het doel van de studie was om de implementatie van een smartphone app, gebaseerd op AI, die huidkanker helpt detecteren te evalueren. De steekproef bestond uit 50 patiënten uit drie huisartspraktijken (13 huisartsen) in Nederland. De patiënten konden de AI-toepassing gebruiken voor de consultatie. De huisartsen onderzochten de patiënten eerst zonder de toepassing, en dan met de toepassing. De primaire uitkomst was het succesvol gebruiken van de app door huisartsen en patiënten. Om de prestatie van de app te beoordelen, werden alle foto's ook door minimum drie dermatologen onafhankelijk beoordeeld, en werden histopathologische gegevens verzameld.	De app had een specificiteit van 80.0% (BI 63.0–91.6) bij het detecteren van goedaardige huidletsels en een specificiteit van 90.9% (BI 55.5–99.8) bij het detecteren van hoog risico huidletsels. 84% van de patiënten gebruikten de app. 90% van de huisartsen gebruikten de app. Huisartsen veranderen hun werkhypothese niet na het gebruik van de app, maar bij vijf patiënten veranderden de huisartsen wel hun behandelplan op basis van de app. 54% van de patiënten van wie de app aangaf dat het huidletsel met hoge waarschijnlijkheid goedaardig was, gaven aan dat ze hun afspraak bij de huisarts zouden annuleren met dit resultaat ($p<.0001$).
Guida et al. (60) O DM en ML Diagnostiek Hartfalen Cardiologie	De eerstelijns databank *'Canadian Primary Care Sentinel Surveillance Network (CPCSSN)'* werd gebruikt in deze studie. De steekproef bestond uit 349 patiënten met hartfalen en 349 patiënten zonder hartfalen. Verschillende variabelen werden geëxtraheerd uit de databank, zoals BMI, bloeddrukwaarden etc. Het doel van de studie was het voorspellen van hartfalen bij de patiënten.	De beste versie van het AI-model had een hoge accuraatheid van 0.89, met een sensitiviteit van 0.89 en een specificiteit van 0.90.

	De dataset werd willekeurig verdeeld in een trainingset (75% van de dataset) en een testset (25% van de dataset) om het AI model te ontwikkelen en te testen.	
Hart et al. (86) O en V ML Doorverwijzing Endometriumcarcinoom Gynaecologie	De steekproef maakte deel uit van de *Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO)*, een prospectieve studie die werd uitgevoerd tussen 1993 en 2001. De steekproef bestond uit 77,254 patiënten zonder kanker (98.8%) en 961 patiënten met kanker (1.2%). De dataset werd verdeeld in een trainingset (70% van de dataset) en een testset (30% van de dataset) om een AI-model te ontwikkelen met als doel het risico op endometriumcarcinoom te voorspellen. De gebruikte variabelen waren: leeftijd, BMI, gewicht, ras, rookgewoonten, diabetes, emfyseem, beroerte, hypertensie, hartziekte, artritis, een andere vorm van kanker, familiegeschiedenis van borst-, eierstok- en baarmoederkanker, operatie aan de eierstokken, leeftijd bij de eerste menstruatie, aantal bevallingen, gebruik van anticonceptie, en leeftijd bij de menopauze. De resultaten van het model werden vergeleken met de resultaten van 15 artsen (gynaecologische oncologen en huisartsen). Deze vergelijking gebeurde bij een subset van 100 patiënten van de oorspronkelijke steekproef. Deze patiënten vulden een vragenlijsten in, en op basis van hun antwoorden classificeerden de artsen de patiënten in twee groepen. Een groep met een verhoogd risico op endometriumcarcinoom en een groep met een verlaagd risico.	Het meest optimale model had een AUC van 0.96, een sensitiviteit van 75.7% en een specificiteit van 98.3%. Dit model was 2.5 keer beter in het identificeren van patiënten met een verhoogd risico op endometriumcarcinoom in vergelijking met het oordeel van artsen, gebaseerd op de antwoorden op een screeningslijst. Verder concluderen de auteurs dat het aantal vals positieve casussen zou halveren met het gebruik van het model.
Herter et al. (45) V ML Behandeling Urineweginfecties Urologie	In deze prospectieve observatiestudie werd een AI toepassing geïmplementeerd in 36 huisartspraktijken gedurende vier maanden (behandelingsgroep, n=2423 patiënten) in Nederland. 29 andere huisartspraktijken uit Nederland (n=3360 patiënten) behoorden tot de controle groep. Gegevens van patiënten met minstens één urineweginfectie werden geëxtraheerd uit de *Nivel Primary Care Database.* De toepassing gebruikt 67 relevante variabelen, op basis van de NHG-standaard, en gaf een voorstel over de antibioticakeuze.	De proportie van succesvolle behandelingen steeg significant in de behandelgroep van 75% naar 80% (z-test=5.47, P<.001). In de controlegroep was er geen significante stijging. Het verschil in succesvol behandelen tussen de behandelgroep en controlegroep tijdens de studie was ook significant (80% voor behandelgroep en 76% voor controlegroep). Er werd geen significant verschil gevonden in voorschrijfgedrag voor en na de start van de studie in beide groepen.

	De proportie van succesvolle behandelingen (=28 dagen na de start van de behandeling werd geen nieuwe antibioticakuur opgestart) werd vergeleken in de behandelingsgroep voor en na de start van het gebruik van de toepassing. De proportie van succesvolle behandelingen tijdens de studie werd ook vergeleken tussen de behandelgroep en de controlegroep. Verder werd het voorschrijfgedrag van de behandelende artsen onderzocht, voor en tijdens de studie, in de behandel- en controlegroep.	
Hill et al. (63) O ML Diagnostiek Voorkamerfibrillatie Cardiologie	De eerstelijns databank *'Clinical Practice Research Datalink'* werd gebruikt in deze retrospectieve studie bij patiënten (>30 jaar) zonder een voorgeschiedenis van voorkamerfibrillatie tussen 2006 en 2016. De steekproef bestond uit 2,994,837 patiënten, 3.2% werd gediagnosticeerd met voorkamerfibrillatie. De casussen werden verdeeld over een trainingset (66,6%) en testset (33.3%). Verschillende AI-modellen werden ontwikkeld om VKF te voorspellen met gebruik van variabelen, waaronder leeftijd, geslacht, BMI, bloeddruk, rookgedrag, medicatie, voorgeschiedenis en comorbiditeiten.	Het meest optimale model had een AUC van 0.827, sensitiviteit van 75%, een specificiteit van 74.9% en een positief voorspellende waarde van 11.5%.
Hill et al. (46) V ML Diagnostiek Voorkamerfibrillatie Cardiologie	De steekproef bestond uit 23745 patiënten van zes verschillende huisartspraktijken in Engeland. De patiënten werden willekeurig verdeeld in twee groepen. Patiënten die de routinezorg kregen en daarnaast werden uitgenodigd voor extra diagnostisch onderzoek (behandelgroep, n=11,849) en patiënten die enkel de routinezorg kregen (controlegroep, n= 11,896). De AI-toepassing werd gebruikt in beide groepen om hoogrisicopatiënten te identificeren. Binnen de behandelgroep, werden patiënten die door een AI-toepassing geïdentificeerd als patiënten met een verhoogde kans op voorkamerfibrillatie (n=944), uitgenodigd voor verder diagnostisch onderzoek. Het gebruikte AI model was reeds ontwikkeld door dezelfde auteurs. De referentiestandaard voor voorkamerfibrillatie werd gesteld door een cardioloog. Verder werden ook analyses uitgevoerd in subgroepen, bijvoorbeeld in de subgroep van patiënten die verder diagnostisch onderzocht werden.	Voorkamerfibrillatie en gerelateerde stoornissen werden gediagnosticeerd in 5.63% en 4.93% van de hoogrisicopatiënten in respectievelijk de controle- en behandelgroep, in vergelijking met in 0.6% van de laagrisicopatiënten (odds ratio 1.15, BI 0.77-1.73, p= 0.486). Van de hoogrisicopatiënten die op een gestandaardiseerde manier diagnostisch onderzocht werden (behandelgroep), kreeg 9.41% een diagnose (odds ratio 2.24, BI 1.31-3.73, p=.003).

Himmelreich et al. (62) V Niet gespecificeerd Diagnostiek Hartritmestoornissen Cardiologie	De steekproef bestond uit 290 patiënten die een ECG kregen bij 11 huisartsen in Nederland. De ECGs werden verwerkt door een *Smartphone App*. De app gaf vervolgens aan welk hartritmestoornissen aanwezig waren (bv. hartritmestoornissen, voorkamerfibrillatie). Dezelfde ECGs werden ook bekeken door experten (huisartsen en cardiologen).	De app had een sensitiviteit van 86% (BI 76%–93%) en een specificiteit van 92% (BI 87%–95%) voor het detecteren van hartritmestoornissen. Voor het detecteren van voorkamerfibrillatie waren de sensitiviteit en specificiteit respectievelijk 97% (BI 85%–100%) and 99% (BI 97%–100%)
Hirosawa et al. (104) V NLP Diagnostiek Algemene geneeskunde	De diagnostische accuraatheid van differentiaal diagnoses voor vaak voorkomende klachten, gegenereerd door *ChatGPT3* werd onderzocht. Er werden 30 casussen uitgevonden met betrekking tot 10 vaak voorkomende klachten in Engeland (bv. hoofdpijn, koorts, buikpijn etc.). Twee beoordelaars (specialisten in algemene geneeskunde) maakten de correcte diagnose en een lijst van vijf differentiaal diagnoses per casus. Volgende prompt werd ingegeven in *ChatGPT3*: “Geef me de top 10 ziekten te verwachten aan de hand van volgende symptomen” vervolgens werden de verschillende casussen ingevoerd. De output werd vergeleken met het oordeel van de twee artsen.	Het aantal correcte diagnoses gemaakt door *ChatGPT-3* binnen de lijst van 10 differentiaal diagnoses was 28/30 (93.3%). Het aantal correcte diagnoses gemaakt door artsen binnen de lijst van 5 differentiaal diagnoses was beter dan de prestatie door *ChatGPT* (98.3% vs. 83.3%, $p = 0.03$). Ook de correcte diagnose was meer accuraat bij artsen (53.3% vs. 93.3%, $p < 0.001$).
Homburg et al. (57) O NLP Diagnostiek en voorspelling COVID-19 Pneumologie	Het *BERT*-model werd aanvankelijk vooraf getraind op Nederlandse taalgegevens en verder afgestemd met behulp van een uitgebreide dataset voor Elektronische Gezondheidsdossiers (EHR), bestaande uit bevestigde COVID-19-huisartsconsultaties en niet-COVID-19-gerelateerde consultaties. De steekproef omvatte 300,359 huisarts consultaties gebaseerd op ICPC-codes die mogelijk verband houden met COVID-19. De database werd verdeeld in 3 sets, waarbij de trainingsset 180,215 consultaties omvatte, waarvan 150,706 niet gerelateerd waren aan COVID-19 en 29,509 gerelateerd waren aan COVID-19. De ontwikkelings- en testsets omvatten elk 60,072 consultaties. De verhouding van niet-COVID-19- tot COVID-19-consultaties voor alle 3 sets was ongeveer 86:14. De dataset werd opgesplitst in een trainings- en validatieset, en de prestaties van het model werden geëvalueerd op een onafhankelijke testset, die diende als de belangrijkste maatstaf voor de effectiviteit ervan in COVID-19-detectie. Om het definitieve model te valideren, werd de prestatie	Het model was zeer accuraat (accuraatheid 0.97, precisie 0.85, positief voorspellende waarde 0.85, sensitiviteit 0.85, specificiteit 0.99). Het validatieonderzoek met de PCR-test toonde een hoge sensitiviteit, maar lage positief voorspellende waarde en specificiteit. Het validatieonderzoek met data uit ziekenhuizen toonde een significante associatie tussen COVID-19 voorspellingen op basis van het model, en COVID-19 gerelateerde hospitalisaties (R^2=.69).

	beoordeeld op 3 manieren. Ten eerste werd externe validatie toegepast op een EHR-dataset uit een ander geografisch gebied in Nederland. Ten tweede werd validatie uitgevoerd met behulp van resultaten van PCR-testgegevens. Ten slotte werd de correlatie tussen voorspelde uitkomsten en COVID-19-gerelateerde ziekenhuisopnames in Nederland beoordeeld, dit voor de periode rond de uitbraak van de pandemie in Nederland (de periode vóór grootschalig testen).	
Hung et al. (93) V ML Diagnostiek Ptose Oftalmologie	In totaal werden 1000 retrospectieve afbeeldingen van het oog van oftalmologie afdelingen gelabeld door drie oftalmologen als zijnde ptose, inclusief echte en pseudo ptose, of als een gezond ooglid. Drie artsen met minstens vijf jaar klinische ervaring, respectievelijk uit spoed, neurologie en huisartsgeneeskunde behoorden tot de niet-oftalmologische groep. Dezelfde testset, bestaande uit 25 gezonde oogleden en 25 ptose, werd aan de groep gegeven om ptose te onderscheiden van gezonde oogleden. Er werd verder geen andere informatie, zoals metingen, de toestand van het andere oog of voorgeschiedenis, gegeven. Bovendien kregen de artsen geen verdere training over de diagnose van ptose. De besluitvorming berustte op de persoonlijke achtergrondkennis van elke arts.	Het model had een sensitiviteit van 0.92 en een specificiteit van 0.88. De niet-oftalmologische artsen behaalden een gemiddelde sensitiviteit van 0.72 en een gemiddelde specificiteit van 0.827. Het AI-model vertoonde een betere prestatie dan de niet-oftalmologen in het identificeren van ptose.
Jain et al. (68) V CV Diagnostiek 120 huidaandoeningen Dermatologie	De steekproef bestond uit 20 huisartsen (en 20 eerstelijnsverpleegkundigen) uit de Verenigde Staten. Beiden onderzochten retrospectief verschillende casussen (n= 1048) met betrekking tot 120 verschillende huidaandoeningen. Door middel van randomisatie werd elke casus beoordeeld met ondersteuning van een AI-toepassing en zonder ondersteuning van een AI-toepassing. De diagnostische accuraatheid werd vergeleken met het oordeel van drie dermatologen. Verder werden ook andere variabelen onderzocht, zoals bijvoorbeeld doorverwijsgedrag, duur van consultatie en vertrouwen in diagnostiek.	Bij de huisartsen was er een stijging in correcte diagnoses (op basis van diagnoses van dermatoloog) van 10% (BI, 8%-11%; P < .001), namelijk van 48% naar 58% bij het gebruik van de AI-toepassing. Doorverwijzingen voor biopsie en doorverwijzingen naar dermatologen daalden met respectievelijk 1% (BI, 0-3%) en 3% (BI, 1%-4%) bij het gebruik van AI. De duur van consultatie steeg met 5 seconden (BI, 0-8) wanneer huisartsen de AI-toepassing gebruikten. Tot slot steeg het zelf gerapporteerde vertrouwen in de diagnostiek ook significant bij het gebruik van de AI-toepassing.
Jammeh et al. (79) O	De dataset bestond uit patiëntengegevens (>65 jaar) van 18 huisartsen in het Verenigd Koninkrijk tussen 2010 en 2012, geëxtraheerd uit een eerstelijns databank. Verzamelde gegevens waren onder andere vergeetachtigheid,	Het beste model had een AUC van 0.869, een sensitiviteit van 84.47% en een specificiteit van 86.67%. 86.06% van de patiënten werd correct geclassificeerd.

ML Diagnostiek Dementie Geriatrie	bloeddruk etc. De steekproef bestond uit 850 patiënten met dementie en 24858 zonder dementie. Verschillende AI-modellen werden ontwikkeld om dementie te identificeren. 9/10 Van de dataset werd gebruikt als trainingset, en 1/10 als testset.	
Judd et al. (94) O en V NLP en ML Diagnostiek Lage rugpijn Fysische geneeskunde	In deze conference paper (preliminair onderzoek) werd data uit de databank *OSCAR EMR* gebruikt. Verschillende gegevens zoals laboresultaten en diagnoses werden gebruikt, maar ook ongestructureerde klinische notities van consultaties in de eerstelijns geneeskunde. De steekproef bestond uit 34 patiënten geselecteerd uit deze databank met klachten over lage rugpijn. Er werden AI modellen, op basis van NLP en ML, ontwikkeld om te differentiëren tussen verschillende soorten rugpijn (bv. gewrichtspijn, zenuwpijn, geen rugpijn etc.) op basis van ongestructureerde notities. De beoordeling van het model werd vergeleken met de beoordeling van een expert aan de hand van de *Clinically Organized Relevant Exam (CORE) Back Tool.* 65% van de dataset werd gebruikt als trainingset, en 35% als testset.	De modellen hadden een zeer hoge accuraatheid, met 100% accuraatheid voor het beste model. Dit werd bereikt in een kleine steekproef. De auteurs geven zelf aan dat dit een pilotstudie is voor een groter vervolgonderzoek.
Kanagasingam et al. (91) O en V CV en ML Diagnostiek Diabetische retinopathie Oftalmologie en endocrinologie	De AI-toepassing werd ontwikkeld op basis van een andere dataset bestaande uit data van verschillende databanken (=30.000 afbeeldingen). Vervolgens werd het ontwikkelde model getest in een steekproef van 193 patiënten met diabetes van vier huisartsenpraktijken uit West-Australië in 2016-2017. Deze patiënten werden gerekruteerd na een bezoek bij de huisarts. 386 Afbeeldingen van hun ogen (gemaakt door verpleegkundigen) werden beoordeeld door een AI-toepassing en door een oftalmoloog voor het identificeren van retinopathie. Er werden slechts twee patiënten gediagnosticeerd met ernstige retinopathie.	De AI-toepassing had een specificiteit van 92% (BI 57%-96%) en een positief voorspellende waarde van 12% (BI 8-18%). De twee patiënten met ernstige retinopathie werden correct geïdentificeerd.
Kennedy et al. (98) O en V ML Diagnostiek	De *Secure Anonymised Information Linkage databank* werd gebruikt, een databank met informatie uit huisartsenpraktijken en ziekenhuizen in het Verenigd Koninkrijk. Patiënten met SA werden geïdentificeerd op basis van de beschikbare data en gelinkt aan controlepatiënten zonder SA. De totale dataset bestond uit 5751 patiënten. De belangrijkste variabelen gebruikt door de	Over het algemeen hadden de AI-modellen een positieve voorspellende waarde van 70%-80% in de testset. De positieve voorspellende waarde was echter maar 0.33%-0.25% in de validatie set.

Spondylitis Ankylosans (SA) Reumatologie	modellen waren musculoskeletale diagnoses, codes en medicatie m.b.t. pijn, laboresultaten, beeldvorming van de rug en uveitis. Verschillende AI-modellen werden ontwikkeld. 10% vande totale dataset werd gebruikt voor het ontwikkeling van de AI-modellen. Waarvan 70% werd gebruikt als trainingset, en 30% als testset. 27% van de totale dataset werd gebruikt om de generaliseerbaarheid van het model te testen. 63% van dataset werd niet geïncludeerd in deze studie omdat deze patiënten voor de leeftijd van 15 jaar gediagnosticeerd werden met SA. De prevalentie van SA was lager in vergelijking met de prevalentie in de gewone populatie.	
Kocks et al. (99) O en V ML Diagnostiek Astma/COPD Pneumologie	Er werd een dataset gemaakt op basis van het elektronisch patientendossier van 400.000 patiënten (>35jaar). Hiervoor werd de Amerikaanse databank *Optum* gebruikt. Uit het patiëntendossier werden 12 parameters gehaald die gebruikt werden om het AI model te trainen. Voor de validatie werd een aparte dataset gecreëerd door 15 Nederlandse artsen, elk 10 nieuwe patiënten met klachten door te laten sturen. Uiteindelijk werden 149 nieuwe casussen verkregen, waar na toepassing van inclusiecriteria, 119 casussen overbleven. Van de 119 casussen werden er 116 voorzien van een diagnose (astma, COPD, astma/COPD overlap en andere) door een expertpanel bestaande uit 3 pneumologen en 4 huisartsen. Een diagnose voor elk van de 116 casussen werd gevraagd aan 180 pneumologen en 180 huisartsen uit verschillende landen. Anderzijds werd een diagnose verkregen door gebruik van de AI-toepassing. Deze resultaten werden vergeleken met de diagnose gesteld door de experten.	De gemiddelde diagnostische accuraatheid was voor de AI-toepassing hoger dan de huisartsen (mediaan verschil 24%, BI 17-29%, P < .0001). Het was ook hoger dan de diagnostische accuraatheid van de pneumologen (mediaan verschil 12%, BI 6-17%, P < .0006). Bij de diagnose van casussen met de specifieke diagnose astma/COPD-overlap was de AI-toepassing inferieur aan zowel huisartsen als pneumologen.
Krentz et al. (67) O ML Behandeling Hyper-cholesterolemie Cardiologie	Voor de dataset werden de elektronische patiëntendossiers opgevraagd van zes Britse huisartsenpraktijken in de periode 1988-2021. In totaal werd een groep van 48.226 patiënten geanalyseerd. Na toepassing van de inclusiecriteria, werden 5630 patiënten die ooit behandeld werden met cholesterolverlagende medicatie overgehouden. Een cohort van 2611 patiënten werd geïdentificeerd die de nodige daling van cholesterol niet haalde.	Van de groep waarvoor de reductiedoelstellingen niet behaald werden, stelde het neurale netwerk in 40% van de gevallen voor om de behandeling voort te zetten. In 46% van de casussen stelde het model voor om de hoeveelheid statines te verhogen. In 5% werd er medicatie toegevoegd en in 9% werd de patiënt doorverwezen. Deze *proof-of-concept* toont het nut aan voor het identificeren van patiënten waar de behandeling dient aangepast te worden, conform de

	Een ML-model (neuraal netwerk) werd ontwikkeld dat de juiste behandeling, die de richtlijnen volgt, voorspelde voor een patiënt. Hiervoor werden een aantal klinische parameters uit het patiëntendossier gebruikt. In het correct volgen van de richtlijn behaalde het neurale netwerk haalde een accuraatheid van 95.9.% Het geoptimaliseerde neurale netwerk werd nadien gebruikt om voor het deel van de patiënten waarvoor geen voldoende reductie in cholesterol gezien werd, aanbevelingen te doen. Twee artsen controleerden de aanbevelingen van het model.	richtlijnen. Geen resultaten over de accuraatheid van het model.
Lee et al. (105) O en V ML Diagnostiek Urinaire flow bij LUTS Urologie	In deze prospectieve studie werden 331 mannelijke patiënten geïncludeerd uit het *Singapore General Hospital* in de periode 2017-2019 (534 patiënten voor exclusie). De patiënten urineerden in een uroflowmetrietoestel, waarbij de geluiden werden opgenomen met een smartphone in de directe omgeving. De referentie standaard werd bepaald door 3 ervaren urologen. Deze studie bestond uit twee delen. Eerst ontwikkelden en evalueerden de auteurs een ML-model (*Audioflow*) om maximum *flow* rate (Qmax), gemiddelde *flow rate* (Qave), en *voided volume* (VV) te bepalen uit de opnames in vergelijking met conventionele uroflowmetrie. Vervolgens werd de mogelijkheid van het model om normale uroflow te onderscheiden van een afwijkende uroflow op basis van de geluidsopname geëvalueerd. Dit werd vergeleken met de resultaten van 6 artsen, waarvan 2 huisartsen en 4 assistenten urologie. Er werd niet beschreven hoeveel huisartsen deelnamen.	De overeenkomst tussen het ML-model en conventionele uroflowmetrie voor Qmax, Qave en VV was respectievelijk 0,77 (BI, 0,72–0,80), 0,85 (BI, 0,82–0,88) en 0,84 (BI, 0,80–0,87). Voor de identificatie van abnormale urineflows behaalde Audioflow een hoge overeenkomst van 83,8% (BI, 77,5–90,1%) met de referentie standaard. Dit was vergelijkbaar met de zes artsen. De AUC van het model was 0,892 (BI, 0,834–0,951), met een sensitiviteit van 87,3% (BI, 76,8–93,7) en specificiteit van 77,5% (BI, 61,1–88,6).
Levkovich et al. (100) V NLP Behandeling Depressie Psychiatrie	Het doel van deze studie was om voorgestelde behandelingen voor milde en ernstige depressie van *ChatGPT* te vergelijken met die van 1249 huisartsen. Hiervoor werden acht casussen van patiënten met depressieve klachten gegenereerd. In de acht casussen werden verschillende eigenschappen systematisch aangepast (bv. geslacht, leeftijd, ernst van symptomen etc.). Vervolgens werd elke casus 10 keer ingevoerd in *ChatGPT-3.5* en *ChatGPT-4* en werd het volgende gevraagd: “wat denk je dat een huisarts zou moeten adviseren in deze situatie?”. De gegenereerde adviezen werden vergeleken met de adviezen van 1249 huisartsen. Verder werd ook onderzocht of er een invloed (bias) was voor sociaaleconomischestatus (SES) en geslacht bij het geven van adviezen.	*ChatGPT-3.5* en *ChatGPT-4* adviseerden bij milde depressie psychotherapie in respectievelijk 95% en 97.5% van de casussen hoewel huisartsen dit slechts deden bij 4.3% van de casussen (p<.001). In het geval van ernstige depressie, adviseerden 44.4% van de huisartsen een combinatie van psychotherapie en farmacologische behandelingen, *ChatGPT* deed dit echter meer frequent, in 72% van de casussen (p<.001). De auteurs gaven hierbij aan dat *ChatGPT* dus vaker de evidence-based richtlijnen voor het behandelen van een depressie volgt in vergelijking met huisartsen. In het geval van farmacologische behandelingen, werden andere

		adviezen gegeven over de soort medicatie door ChatGPT in vergelijking met huisartsen. De adviezen van *ChatGPT* werden niet beïnvloed door geslacht en/of SES in tegenstelling tot de adviezen van huisartsen.
Liu et al. (69) O en V ML Diagnostiek 26 huidaandoeningen Dermatologie	19.870 casussen (afbeeldingen en klinische gegevens) werden verzameld tussen 2010 en 2018 uit een dermatologiepraktijk, actief op 17 verschillende locaties in 2 staten van de Verenigde Staten. De patiënten geregistreerd tussen 2010-2017 werden gebruikt als trainingset (n=16,114) en de patiënten geregistreerd tussen 2017-2018 gebruikt als validatieset (n=3,756). De validatieset werd nog verder verfijnd in een testset (n=963), die werd gebruikt voor de vergelijking met de artsen. Deze testset had in verhouding een grotere prevalentie van zeldzamere aandoeningen. Het model werd eerst ontwikkeld met de trainingset en validatieset. Nadien werden de resultaten van het model geëvalueerd op de testset. Tot slot werden de resultaten van het model op de testset vergeleken met de resultaten van 18 zorgverleners, waarvan 6 huisartsen, 6 dermatologen en 6 verpleegkundigen (nurse practitioners). De diagnostiek van 3 dermatologen werd gebruikt als norm.	Het model behaalde een top-1 accuraatheid van 0.66, vergeleken met 0.63 voor de dermatologen, 0.44 voor de huisartsen en 0.40 voor de verpleegkundigen. Het artikel besluit dat het model gelijkwaardig presteerde met de dermatologen. Verder was de top-3 accuraatheid van het model hoger, 0.90. Dit vergeleken met 0.75 voor de dermatologen, 0.60 voor de huisartsen en 0.55 voor de verpleegkundigen.
Livingstone et al. (106) O en V CV en ML Diagnostiek Otoscopie NKO	Er werden prospectief otoscopische beelden verzameld in het ziekenhuis *Alberta Health Services facilities* in Calgary. Zowel pathologische als gezonde beelden werden verzameld. Verder werd de dataset van Google Images doorzocht op otoscopische beelden. Daarnaast werden er beelden verzameld uit onder andere online datasets, tekstboeken en andere databanken. De referentie diagnose werd bepaald door een ervaren NKO-arts en een assistent. Een steekproef van 1336 verschillende afbeeldingen werd weerhouden. Er werd een trainingset en testset gecreëerd. Na training van het ML-model werden de resultaten vergeleken met de resultaten van artsen, op 89 nieuwe beelden. Er namen 10 artsen deel, waarvan 2 ervaren huisartsen.	Het model had een gemiddelde positief voorspellende waarde van 90,9% met een sensitiviteit van 86,1% in de testset. Bij de afbeelding (n=89) uit de vergelijking met artsen, maakte het model 79 correcte diagnoses met een accuraatheid van 88,7%. De gemiddelde accuraatheid van de artsen was hiervoor 58,9%.
Lucius et al. (71) O en V	Deze studie gebruikte afbeeldingen uit de *HAM10000-dataset* via het *International Skin Imaging Collaboration (ISIC)* archief. De totale steekproef van	In het eerste deel hadden de 22 huisartsen een gemiddelde accuraatheid van 27,74%. Het minst goede ML model haalde een accuraatheid van 78.4% op dezelfde afbeeldingen.

ML Diagnostiek 7 huidletsels Dermatologie	10,015 dermatoscopische afbeeldingen werd verdeeld in een trainingset (8313, 83%) en een testset (1702, 17%). 8 verschillende ML-modellen(*deep neural networks*) werden eerst getraind en vervolgens getest. Er werd nadien een vergelijking gemaakt tussen de resultaten van de modellen en met de resultaten van 22 huisartsen uit Buenos Aires, Argentinië. De artsen kregen 163 beelden te zien uit dezelfde dataset en moesten die diagnosticeren. Hiervoor werden 19 huisartsen eerst gevraagd om 35 afbeeldingen te beoordelen. In een tweede stap hadden artsen toegang tot de voorspellingen van de modellen, waarbij dezelfde groep elke afbeelding classificeerde op basis van zowel hun criteria als het resultaat van de modellen. Hiervoor werd een nieuwe set van 35 afbeeldingen aangeboden.	In het tweede deel, waar 19 huisartsen 35 afbeeldingen moesten beoordelen, was de accuraatheid 17,29%. In dezelfde steekproef behaalde het beste ML-model een accuraatheid van 77,14%, wat beter was dan de artsen. Toen huisartsen de mogelijkheid kregen om het resultaat van het model per afbeelding te bekijken, steeg de accuraatheid naar 42,42%. Een toename van 25,13%.
Maclagan et al. (80) O NLP en ML Diagnostiek Dementie Geriatrie	De steekproef is afkomstig uit de *Electronic Medical Records Primary Care databank.* Deze databank omvat elektronisch patiëntendossier van 350 huisartsen en ongeveer 400,000 bijhorende patiënten. 526 patiënten met dementie en 44,148 controlepatiënten zonder dementie werden geïncludeerd. Patiënten werden geïncludeerd als er gegevens waren minimaal 2 jaar voor de diagnose en 2 jaar na de diagnose. Een lijst van kenmerken van dementie werd samengesteld op basis van literatuur en de voorstellen van 3 artsen (1 neuroloog, 1 geriatrische psychiater en 1 huisarts). Deze kenmerken werden gebundeld en bruikbaar gemaakt voor het AI model. De klinische notities en de kenmerken werden gebruikt door het model om patiënten met dementie te identificeren uit de steekproef.	Indien alle gegevens werden gebruikt, had het beste ML-model een accuraatheid van 77.2%, met een sensitiviteit van 71.5% en specificiteit van 99.8%. Dit model gebruikte een combinatie van de kenmerken en klinische notities. Echter indien men enkel de gegevens gebruikte voor de datum van diagnose, daalde de prestatie van het model aanzienlijk met een accuraatheid van 14.4%, sensitiviteit van 8.3% en specificiteit van 99.9%.
Malhotra et al. (84) O ML Diagnostiek Pancreascarcinoom Gastroenterologie	De dataset kwam uit *The English National Cancer Registry (CR)* die verbonden is met *Clinical Practice Research Datalink (CPRD) databank,* 1.139 patiënten tussen leeftijd van 15-99 jaar, die gediagnosticeerd waren met pancreascarcinoom tussen 2005 en 2009 werden geïncludeerd. Elke patiënt werd gekoppeld aan vier controles met niet-pancreatische maligniteit. De maligniteit van de controles mocht geen overlappende symptomen of risicofactor hebben met pancreascarcinoom, hierdoor werden carcinomen uitgesloten van lip, mond & keel, verteringstelsel, luchtwegen en intrathoracale organen, borstcarcinoom en gynaecologische carcinomen. De gegevens van 24 maanden voor de diagnose werden gebruikt om 57 variabelen te selecteren,	De AI-toepassing kon 41.3% van de patiënten jonger dan 60 jaar identificeren als hoog risico op het ontwikkelen van pancreascarcinoom binnen 20 maanden. Dit met een AUC van 66%, sensitiviteit van 72,5% en specificiteit van 59%. Verder kon de toepassing 43.2% van de patiënten ouder dan 60 jaar identificeren als hoog risico op het ontwikkelen van pancreascarcinoom binnen de 17 maanden. Dit met een AUC van 61%, sensitiviteit van 65% en specificiteit van 57%.

	die werden gebruikt als data voor het model te trainen en evalueren. Het ML-model werd getraind op 75% van de data en geëvalueerd op de overige 25%.	
Muñoz-López et al. (107) V ML Diagnostiek Huidlaesies Dermatologie	De steekproef bestond uit 340 afbeeldingen met 87 verschillende diagnoses afkomstig van 281 patiënten. Deze data werd verzameld in een dermatologische praktijk uit Chili van maart t.e.m. april 2020. Er kon meer dan 1 aandoening aanwezig zijn per patiënt. Casussen werden geëxcludeerd indien er geen afbeelding gemaakt werd of van slechte kwaliteit was, indien er geen afwijking op de afbeelding te zien was of indien er geen verder onderzoek werd gedaan ondanks advies van de dermatoloog. De referentie standaard bestond uit diagnostiek na vervolg consultatie, huidbiopsie of door overeenkomst met een panel van zes dermatologen. Het ML-model was reeds getraind door één van de auteurs. De studie evalueerde de resultaten van het model in vergelijking met de referentiestandaard. Verder werd er vergeleken met de resultaten van 9 artsen, waaronder 3 huisartsen, 3 dermatologen en 3 dermatologie assistenten.	De top-1 accuraatheid van het model (41,2%) was lager dan die van dermatologen (60,1%), assistenten (57,8%) en huisartsen (49,3%) ($p<0.05$). Wanneer men enkel de analyse bekeek van de diagnosen waarop het model explicit was getraind, was de top-1 accuraatheid (47.6%) vergelijkbaar met dermatologen (49,7%), assistenten (47,7%) maar beter dan huisartsen(39,7%, $p = 0,049$).
Nadarajah et al. (64) O ML Diagnostiek Voorkamerfibrillatie Cardiologie	De *databank Clinical Practice Research datalink* (CPRD) werd gebruikt; Er werden 2.081.139 patiënten geïncludeerd tussen 1998 en 2018. Van deze groep werd er VKF vastgesteld bij 7386 personen(0,4%). Het doel was het risico bepalen op VKF in 6 maand. 80% van de dataset behoorde tot de trainingset, en 20% tot de testset. Het ML- model, genaamd *FIND-AF*, gebruikte verschillende variabelen waaronder leeftijd, geslacht, etniciteit en comorbiditeiten (72 comorbiditeit gecodeerd als binaire variabelen; aan- of afwezig). Bloedwaarden werden niet gebruikt. Na training werd het resultaat van model op de testset vergeleken met het resultaat van de CHA_2DS_2-VASc score en C_2HEST score.	*FIND-AF* had een AUC van 0.824 (BI 0.814-0.834), sensitiviteit van 0.781 (BI 0.731-0.829), specificiteit van 0.731 (BI 0.693-0.771), positief voorspellende waarde van 2.5% (BI 2.3-2.7), negatief voorspellende waarde van 99.8% (BI 99.8-99.8). Dit in vergelijking met CHA_2DS_2-VASc met een AUC van 0.784 (BI 0.773-0.794), sensitiviteit van 0.847 (BI 0.829-0.866), specificiteit van 0.611 (BI 0.608-0.612), positief voorspellende waarde van 2.2% (BI 2.1-2.3), negatief voorspellende waarde van 99.8% (BI 99.8-99.8) en C_2HEST met een AUC van 0.757 (BI 0.744-0.770), sensitiviteit van 0.642 (BI 0.619-0.791), specificiteit van 0.790 (BI 0.622-0.792), positief voorspellende waarde van 2.0% (BI 1.5-2.2), negatief voorspellende waarde van 99.7% (BI 99.7-99.8).

		De auteurs concludeerden dat het model beter presteerde dan de twee scores, met nadruk op de het verschil in positief voorspellende waarde.
Nemlander et al. (85) O M-L Diagnostiek Ongemetastaseerde colorectaal-carcinoom Gastroenterologie	De dataset komt uit een grootschalig bevolkingsonderzoek waarbij gegevens zijn verzameld van alle patiënten ouder dan 18 in de regio *Vastra Gotaland* die in 2011 werden gediagnosticeerd met ongemetastaseerde colorectaalcarcinoom. De deelnemers moesten in leven zijn op het moment van de diagnose, in het jaar voorafgaand aan de diagnose een huisarts van VGR hebben geraadpleegd, en gedurende de afgelopen 20 jaar niet zijn gediagnosticeerd met kanker. Elke patiënt werd gematcht op leeftijd, geslacht en huisartsenpraktijk met maximaal vier controles. De analyse omvatte gegevens van 542 patiënten met ongemetastaseerde colorectaalcarcinoom en hun 2.139 gematchte controles. De gecodeerde diagnoses van de consultaties werden gebruikt. 75% van de patiënten (n=407) met overeenkomende controles (n=1606) werden willekeurig geselecteerd voor het trainen van het AI-model, terwijl de overige 25% van de patiënten (n=135) met bijpassende controles (n=533) werden gebruikt voor het evalueren van de prestaties. Het AI-model werd toegepast op de testdataset om de individuele kans op ongemetastaseerde colorectaalcarcinoom voor elke patiënt te berekenen. De grenswaarde voor het classificeren van patiënten met ongemetastaseerde colorectaalcarcinoom werd bepaald als de waarde die de som van gevoeligheid en specificiteit maximaliseerde. Patiënten in de testset werden dan geclassificeerd als het hebben van ongemetastaseerde colorectaalcarcinoom als de individuele kansen, verkregen uit het AI-model, groter waren dan deze grenswaarde, en anders geclassificeerd als het niet hebben van ongemetastaseerde colorectaalcarcinoom.	Het AI-model had een sensitiviteit van 0.733, een specificiteit van 0.835 en een AUC van 0.832 (BI 0.790;0.874).
Oude Nijeweme-d'Hollosy et al. (50) O en V	Het doel van deze studie was om een AI-model te ontwikkelen en testen, dat patiënten met lage rugpijn zelf kunnen gebruiken om advies te krijgen rond doorverwijzing . De trainingset (70%) en validatieset/testset (30%) maakten gebruik van data verzameld voor een eerdere studie met eerstelijnszorgverleners om drie	De accuraatheid van de drie modellen (decision tree, the random forest and boosted tree) was respectievelijk 70%, 69% en 72%.

ML Doorverwijzing Lage rugpijn Fysische geneeskunde	AI-modellen te ontwikkelen. De steekproef omvatte 1288 fictieve casussen van lage rugpijn, beoordeeld door 63 kinesisten en huisartsen op advies.De 3 modellen gaven een advies voor doorverwijzing voor de casussen. Dit advies werd vergeleken met het advies van de zorgverleners. De onafhankelijke testset bestond uit een prospectieve studie met 38 patiënten van wie 33 onderzocht werden door een kinesist en 5 door een huisarts. De zorgverleners gaven een advies na het onderzoeken van de patiënten. Dit advies werd vergeleken met het advies van het AI-model.	In de onafhankelijke testset was de accuraatheid respectievelijk 71%, 53% en 71%. Voor het meest optimale model (boosted tree) was de sensitiviteit, met het advies van huisartsen als norm, 25% en de specificiteit 79%.
Ozturk et al. (52) O ML Voorspelling Hypertensie bij Type 2 Diabetes (T2D) Endocrinologie	De gebruikte dataset was afkomstig van de *Connected Bradford databank*. Dit is een databank met eerstelijnsinstellingen uit het Verenigd Koninkrijk. Er werden 467000 patiënten met T2D geïncludeerd met bijhorende data. Vervolgens werd een steekproef van 43000 patiënten willekeurig geselecteerd. Aan de hand van herhaalde bloeddrukmetingen werden patiënten geclassificeerd als hoog risico op hypertensie (bloeddrukwaarden meermaals >/= 140/90) en laag risico (<140/90). Er werden AI-modellen ontwikkeld om het risico op hypertensie te voorspellen. De gebruikte variabelen waren bijvoorbeeld BMI, aantal neutrofielen en cholesterol. 80% van de dataset werd gebruikt als trainingset en 20% als testset.	Het meest optimale model had een accuraatheid van 0,9525 en kappa waarde van 0,2183 om T2D patiënten met een hoog risico op hypertensie te identificeren.
Pareja-Ríos et al. (90) O en V ML Diagnostiek Diabetische retinopathie Endocrinologie en oftalmologie	Een databank van 475,330 afbeeldingen, verzameld door huisartsen, oogartsen, in de periode 2007-2017 werd gebruikt. Er werd een foto gemaakt van elk oog tijdens de consultatie (237,665 consultaties), een patiënt kon meermaals geïncludeerd worden indien hij meermaals op consultatie kwam in bovenstaande periode. Van het totale aantal bezoeken werden 149.987 (63,1%) uitsluitend beoordeeld door huisartsen, 33.591 (14,1%) uitsluitend door oogartsen (twijfelachtig of doorverwezen zonder diagnose), en 54.087 (22,8%) door zowel huisartsen als oogartsen. Enkel de beelden uit de laatste twee categorieën werden geïncludeerd in de training (80%) en validatieset (20%) voor deze studie. Het oordeel van een oftalmoloog (bv. geen tekenen van retinopathie, milde retinopathie etc.) werd gebruikt als de norm. De finale dataset bestond uit 11330 afbeeldingen zonder retinopathie en 4529 met minstens milde retinopathie.	Het AI-model identificeerde meer dan 95% van de patiënten met milde tot ernstige diabetische retinopathie. In vergelijking met de beoordeling door huisartsen waren er 70% minder vals positieven. Het AI-model kon 1258 patiënten (=7.9% van het totale aantal patiënten gediagnosticeerd met retinopathie door huisartsen) identificeren voor ze werden geïdentificeerd door huisartsen.

	Vervolgens werden de beelden die niet gebruikt werden in de voorgaande fase gebruikt om het AI-model op te testen (221806 afbeeldingen). Het oordeel van de oftalmoloog was opnieuw de norm. De afbeeldingen werden door de AI-toepassing geclassificeerd als pathologisch of niet pathologisch. De resultaten van het AI-model werden vergeleken met de resultaten van huisartsen en oftalmologen.	
Parente et al. (89) O en V ML Diagnostiek Urineweginfecties Urologie	De gebruikte dataset bestond uit klinische informatie van 80.378 patiënten in een spoedsetting en 472 patiënten in de setting van de eerstelijnsgezondheidszorg en bijhorende resultaten van de aanwezigheid van urineweginfectie. De klinische informatie bestond uit demografische gegevens zoals leeftijd en geslacht, laboresultaten zoals een leukocytentelling, symptomen zoals buikpijn en medische geschiedenis. De eerste groep werd gebruikt om de AI modellen te trainen en (intern) te testen, de tweede gold als externe testset. Verschillende machine learning modellen werden ontwikkeld om een positieve urinekweek te voorspellen. De machine learning modellen waren neurale netwerken, random forests en extreme gradient boosting. De AI-toepassing werd *'NoMicro'* genoemd. De prestatie van dit model werd vergeleken met die van een andere AI-toepassing *'NeedMicro'*, die informatie gebruikt die niet beschikbaar is in de eerstelijnszorg.	De prestatie van *NeedMicro* en *NoMicro* was gelijkaardig. Op de interne testset gaf *NoMicro* een AUC van 0.862 (BI: 0.856-0.869) versus 0.877 (BI: 0.871-0.884) voor *NeedMicro*. Op de externe testset was de AUC voor *NoMicro* ook hoog: 0.850 [BI: 0.808-0.889].
Phillips et al. (72) O en V ML Diagnostiek Melanoma Dermatologie	De gebruikte dataset bestond uit 7102 unieke afbeeldingen van gepigmenteerde huidletsels. 24% van de letsels waren histologisch bevestigde melanomen. Er werden verschillende databanken gebruikt om deze afbeeldingen te bekomen, waaronder de *PH2 dataset, Interactive Atlas of Dermoscopy,* en *ISIC archive.* Deze studie had als doel de nauwkeurigheid te beoordelen van het DERM model (*Deep Ensemble for Recognition of Melanoma*) om melanoom te identificeren op basis van beelden van gepigmenteerde huidletsels. Dit werd gevolgd door een meta-analyse waarbij de diagnostiek van artsen(zowel specialisten als huisarts) werd onderzocht. Deze resultaten werden vergeleken met de resultaten van DERM. DERM werd getraind en getest met de beschreven steekproef. De meta-analyse was gebaseerd op studies met een	DERM had een AUC van 0,93(BI : 0.92-94) met een sensitiviteit van 85% en specificiteit van 85,3%. Uit de meta-analyse gebaseerd op 10 studies, concludeerden de auteurs dat de huisartsen een AUC van 0,83(BI: 0,79-0,86) bereikten met een sensitiviteit van 79.9% en een specificiteit van 70.9%.

	totaal van 32,226 gepigmenteerde huidletsels waarvan 3277 histologisch bevestigde melanomen.	
Pyne et al. (108) O NLP Informatie extractie ICPC-2 codes Algemeen	De '*One in a Million*' (OIAM) databank werd gebruikt voor deze studie. Het bestond uit 300 video- en audio-opnames en woord voor woord verslagen van consultaties die plaatsvonden in 12 verschillende huisartspraktijken te Bristol in het Verenigd Koninkrijk. Van deze databank waren er 239 consultaties die notities bevatten en voorzien waren van ICPC-2 codes. Deze 239 werden gebruikt als steekproef in deze studie. Daarnaast was er nog een dataset, gebaseerd op de databank *National Institute for Health and Care Clinical Knowledge Summaries (NICE CKS)*. Deze dataset bestond uit tekst met samenvattingen van verschillende medische topics. Een steekproef van deze dataset werd verkregen, door de topics te koppelen aan ICPC-2 codes. Er werden verschillende AI-modellen ontwikkeld, met als doel om de consultaties correct te koppelen aan een ICPC-2 code op basis van woord-voor-woord transcripten van consultaties. In de eerste methode werd uitsluitend de steekproef uit OIAM (80% in de training set, 20% in de testset) gebruikt. In het tweede methode werd de steekproef uit NICE CKS met de definities van ICPC-2 codes gebruikt om de modellen te trainen. Verschillende modellen werden uitgetest, die één of meerdere ICPC-2 codes koppelden aan de transcriptie van een consultatie.	De beste resultaten werden bereikt met het BERT-model, getraind volgens de tweede methode. Het behaalde een accuraatheid van 0.51 en een *recall* van 56% en *precisie* van 55%. Wanneer men de woorden van de patiënt excludeerde uit het transcript en het zich enkel baseerde op de woorden van de arts, daalde de accuraatheid.
Ristanoski et al. (87) O ML Diagnostiek Longkanker Pneumologie	De *NPS MedicineInsight databank* is een databank met gegevens van meer dan 500 huisartspraktijken en 5000 huisartsen. Voor deze steekproef werd uit deze databank een longkanker cohorte gevormd, alsook een controle cohorte. De steekproefgrootte van de verschillende groepen werd niet gespecificeerd. De studie gebruikte 5 labowaarden: het aantal bloedplaatjes, de ratio hemoglobine/hematocriet (MCHC), het gemiddelde volume van de rode bloedcellen (MCV), de gemiddelde hoeveelheid hemoglobine per rode bloedcel (MCH) en de variatie in de grootte van de rode bloedcellen (RDW). AI-modellen werden ontwikkeld om een kankerdiagnose te voorspellen op basis van de aanwezigheid van resultaten buiten het normale bereik voor elk van de vijf	Het AI-model om de diagnose van longkanker op drie maanden te voorspellen had een goede nauwkeurigheid (7-8 van de 10 waren correct geclassificeerde patiënten). Men kon ook vergelijkbare nauwkeurigheid bieden met de voorspelling op 6 maanden. De individuele modellen, wanneer ze gecombineerd werden in een *'ensemble'* (vanaf dat één van de verschillende modellen longkanker voorspelt, voorspelt het ensemble longkanker), presteerden goed voor de voorspelling van 3 maanden vooruit, maar niet zo goed voor de voorspelling van 6 maanden vooruit. Het *Stack*-model, dat rekening houdt met de verschillende voorspellingen, leek dit

	bloedonderzoeken (in de perioden van 24–18, 18–12, 12–6 en 6–3 maanden voor de datum van de kankerdiagnose).	probleem niet te hebben. De waarden van *TPR*, *TNR*, *PPV* en *NPV* waren 0,65-0,75 voor de verschillende modellen.
Rushlow et al. (47) V ML Diagnostiek Hartritmestoorniss en en hartfalen Cardiologie	Deze studie bestond uit 358 clinici in 48 Amerikaanse centra, waarvan er 181 willekeurig werden geselecteerd voor de interventiegroep. In de controlegroep zaten 112 artsen, de andere beroepen waren verpleegkundigen en praktijkassistenten. Van alle zorgverleners waren er 118 te situeren in de eerstelijnsgezondheidszorg. Het exact aantal huisartsen werd niet gespecificeerd. Op basis van de patiëntendossiers werden patiënten geïncludeerd die 18 jaar of ouder waren en een ECG lieten afnemen. Na exclusie van de patiënten met een gekende lage ejectiefractie of een diagnose van hartfalen voor de afname van het ECG, werden er 11,573 patiënten geïncludeerd. Het doel van de studie was om de impact te meten van het gebruik van een AI-model op de doorverwijzing voor het diagnosticeren van een lage ejectiefractie. Het AI-model kon op basis van een ECG waarschuwen voor een risico op ventriculaire dysfunctie. De mate van gebruik werd gedefinieerd door het aantal doorverwijzingen voor extra testen voor AI-positieve ECG's te vergelijken met het totaal aantal AI-positieve ECG's. Dit per arts. Zo kreeg je een groep artsen die er veel rekening mee hield (*high adopters*=bovenste kwartiel) en minder rekening mee hield (*low adopters*=3 onderste kwartielen).	Bij de patiënten met een positieve AI-ECG, waren de high adopters dubbel zo waarschijnlijk om de patiënt met lage ejectiefractie te diagnosticeren in vergelijking met low adopters. *High adopters* (33.9%) vs *low adopters*, (16.9%); odds ratio, 1.62; BI, 1.21 to 2.17) (33.9% vs 16.3%, P<.001).
Sekelj et al. (65) V ML Diagnostiek Voorkamerfibrillatie (VKF) Cardiologie	De data komt uit de *DISCOVER* studie, en bestond uit patiënten uit de algemene populatie van Londen. Patiënten werden geïncludeerd indien ze ouder dan 30 jaar waren in de periode tussen 1 januari 2006 en 31 december 2016 en ze geen voorgeschiedenis van voorkamerfibrillatie in de voorafgaande vijf jaar kenden. Alleen patiënten waarvan lengte, gewicht, BMI, systolische bloeddruk en diastolische bloeddruk gekend waren in een welbepaalde periode van 12 maanden kwamen in aanmerking voor opname in de studie. Van de 2,542,732 patiënten in de *DISCOVER*-database voldeed 604.135 patiënten (23,8%) aan de criteria en deze werden opgenomen in de studie. Tijdens de *follow-up* kregen 17.880 (3,0%) patiënten de diagnose van VKF. Een reeds ontwikkeld model werd geëvalueerd aan de hand van de steekproef. De	De AUC was 0,87, vergeleken met 0,83 bij de ontwikkeling van het algoritme. Bij patiënten ouder dan 30 jaar identificeerde het algoritme 99,1% van de patiënten die geen VKF hadden correct (NPV) en 75,0% van de gevallen met VKF (sensitiviteit). Bij patiënten ouder dan 65 jaar (n = 117.965) was de NPV 96,7% en de sensitiviteit 91,8%..

	resultaten van deze studie werden vergeleken met de resultaten van de studie waarin het model was ontwikkeld.	
Soerensen et al. (82) O ML Diagnostiek Maligniteiten Oncologie	Een batterij van labotesten met 25 voorgestelde bloedonderzoeken werd uitgewerkt en voorgesteld aan huisartsen. Dit om te gebruiken bij patiënten met aspecifieke symptomen met als doel patiënten te identificeren met een verhoogd risico op kanker. Deze testbatterijen werden verzameld van 2011 tot 2020 in Denemarken. 6592 resultaten werden weerhouden na toepassing van de inclusiecriteria. ML-modellen werden geëvalueerd op hun vermogen om patiënten te identificeren die een verhoogd risico hadden op een kankerdiagnose binnen de 90 dagen op basis van de batterij van labotesten.	De onderzochte modellen gaven een score tussen 0 en 100 aan de patiënt die het risico op een kankerdiagnose binnen de 90 dagen kwantificeerde. Als er een threshold toegepast werd op deze score, werd op de validatie en testset een AUC van respectievelijk 0.86 en 0.79 behaald. Zo had bij een score gelijk aan 3 de patiënt 10.7% kans op een diagnose van kanker binnen de 90 dagen. Dit met een sensitiviteit van 0.82 en specificiteit van 0.55.
Suárez-Araujo et al. (81) O en V ML Diagnostiek Milde cognitieve achterstand Geriatrie	Uit de *AIzheimer's Disease Neuroimaging Initiative* werd, na toepassing van de inclusiecriteria, een steekproef verkregen, die bestond uit 128 proefpersonen met milde cognitieve achterstand en 203 controles. Een neuraal netwerk dat milde cognitieve achterstand voorspelt, kon gebruik maken van de demografische kenmerken van de proefpersonen, het opleidingsniveau en enkele testresultaten, namelijk de scores van de *Mini Mental Status Examination (MMSE),* de *Functional Assessment Questionnaire (FAQ)* en de *Geriatric Depression Scale.* Een vergelijking van de behaalde resultaten werd gemaakt met die van medische experts, waaronder twee huisartsen, een neuroloog en een geriater.	Het beste AI-model behaalde een AUC van 0.952, een sensitiviteit van 0.9 en een specificiteit van 0.848 en baseerde zich hiervoor op MMSE, FAQ en leeftijd. Artsen hadden een sensitiviteit van 0.467 en een specificiteit van 0.913. Er was veel variabiliteit tussen de artsen in de behaalde resultaten. Over het algemeen behaalden huisartsen de laagste sensitiviteit en specificiteit.
Surendra et al. (61) O ML Diagnostiek Hartfalen Cardiologie	In dit onderzoek werden in totaal 5299 deelnemers van de *Hamburg City Health* studie opgenomen, waarvan de ECG's, die minstens 30 seconden duurden, gebruikt werden. Hiervan waren 4981 patiënten gezond en er waren 318 patiënten met hartfalen. Een neuraal netwerk werd gebruikt om de data te analyseren en de patiënten te classificeren met hartfalen of die gezond zijn. Er werden drie AI-modellen ontwikkeld en vergeleken.	Er werd een AUC van 0,75, een sensitiviteit van 0,67 en een specificiteit van 0,69 gevonden voor het neuraal netwerk; een AUC van 0,77, een sensitiviteit van 0,63 en een specificiteit van 0,76 voor logistieke regressie; AUC 0,79, een sensitiviteit van 0,67 en een specificiteit van 0,72 voor de random forest-classifier.
Topaz et al. (58)	Deze studie werd uitgevoerd in twee gezondheidscentra in Philadelphia. In de studie werd therapietrouwheid bij astma (nemen van medicatie) onderzocht bij	De nauwkeurigheid van spraak-naar-tekst-transcriptie, gekwantificeerd volgens de ROUGE-methode, was hoog

O en V ML Voorspelling Nemen van astma medicatie en kwaliteit van *shared decision making* tijdens consultatie. Pneumologie	patiënten van zorgverleners (huisartsen of verpleegkundigen). De steekproef bestond uit 80 patiënten met ongecontroleerd astma. De tien zorgverleners werden ingedeeld in 2 gelijke groepen. Elke groep onderzocht 40 patiënten. De eerste groep gebruikte een gestandaardiseerde interventie en de tweede groep had een gesprek over gezonde levensstijl maar zonder richtlijn. Deze gesprekken werden opgenomen. Alle patiënten vulden nadien de SDM-Q-9 in, een vragenlijst ter evaluatie van de subjectieve maat van SDM (*shared decision making*). De MARS-10 vragenlijst, een vragenlijst ter evaluatie van de therapietrouw bij astma, werd ook afgenomen. Deze werd ingevuld door de patiënt op 4 momenten, voor het contact met de zorgverleners en op 3 momenten na het contact. De studie onderzocht eerst in welke mate een AI toepassing, Amazon Transcribe, de opgenomen gesprekken accuraat kon neerschrijven (*speech to text*). Verder werd een AI toepassing ontwikkeld die de mate van SDM tracht te voorspellen na het gesprek. De mate van SDM werd vastgelegd volgens SDM-Q-9, een vragenlijst die de subjectieve mate van SDM bij de patiënt evalueert. En tot slot werd de AI toepassing geëvalueerd op de mogelijkheid tot het voorspellen van therapietrouw.	(F-score = 0.9, sensitiviteit = 0.90, positief voorspellende waarde = 0.91). ML-modellen behaalden goede voorspellende waarden voor SDM (de hoogste F-score = sensitiviteit = positieve predictieve waarde = .88 voor *Naive Bayes*) en therapietrouw aan inhalatiecorticosteroïden (de hoogste F-score = .87 voor *Support Vector Machines*, met een sensitiviteit = .88 en positief voorspellende waarde = .87).
Tschandl et al. (70) O en V CV en ML Diagnostiek Gepigmenteerde huid laesies Dermatologie	De dataset voor de ontwikkeling van de toepassing werd samengesteld met dermatoscopische afbeeldingen van de *Dermatologic Imaging Research Group* (ViDIR) van de afdeling dermatologie van de medische universiteit van Wenen, Oostenrijk en van de praktijk van *Cliff Rosendahl* uit Queensland, Australië. 11,210 afbeeldingen werden opgesplitst in een trainingset (n=10,015) en een testset (n=1195). De steekproef voor de evaluatie van het model werd verkregen uit verschillende instanties, waaronder opnieuw de afdeling dermatologie van de medische universiteit van Wenen en de praktijk van *Cliff Rosendahl* uit Queensland maar verder ook van verschillende centra uit Turkije, Nieuw Zeeland, Zweden, en Argentinië. De gebruikte steekproef voor de evaluatie omvatte 1511 dermatoscopische beelden. Verschillende AI-toepassingen werden ontwikkeld in het kader van een competitie tussen 77 deelnemende teams. De steekproef voor de evaluatie werd beschikbaar gemaakt voor de top 3 modellen van deze competitie. De resultaten van deze modellen werden vergeleken met de deelnemende artsen. Er waren 511 deelnemende artsen, waaronder 83 huisartsen, 118 assistenten	De modellen gaven gemiddeld 2,01 (95% betrouwbaarheidsinterval 1,97 tot 2,04; p<0,0001) extra correcte diagnoses in vergelijking met de volledige groep van dermatologen, dermatologie-assistenten en huisartsen, (op 30 afbeeldingen gaven ze respectievelijk 17,91 (SD 3,42) vs. 19,92 (4,27) correcte diagnoses). De groep van 27 artsen met meer dan 10 jaar ervaring gaf gemiddeld 18,78 (SD 3,15) correcte diagnoses, terwijl de top drie AI-modellen gemiddeld 25,43 (SD 1,95) correcte diagnoses stelden (gemiddeld verschil 6,65, 95% betrouwbaarheidsinterval 6,06–7,25; p<0,0001). Bij het tonen van afbeeldingen die verzameld waren in centra die geen afbeeldingen leverden voor in de trainingset zakte de accuraatheid minder bij de top drie algoritmen dan bij artsen (artsen presteerden 11,4% minder, 95% betrouwbaarheidsinterval 9,9–12,9 vs. 3,6%, 0,8–6,3; p<0,0001). Echter, in steekproeven met enkel afbeeldingen

	dermatologie en 283 dermatologen. De artsen kregen via een online toepassing telkens 30 afbeeldingen ter evaluatie. Een arts kon meermaals deelnemen met telkens een nieuwe 30 afbeeldingen.	van kwaadaardige huidletsels scoorden de artsen beter dan de top drie AI-modellen die respectievelijk 0.54, 0.76 en 0.33 meer fouten maakten (p<0.0001).
van Mens et al. (55) O ML Voorspelling Suïcide Psychiatrie	De dataset bestond uit gegevens van de *Nivel Primary Care Database*. Deze databank bestond uit data meer dan 1,5 miljoen patiënten uit de eerstelijnszorg. Patiënten die een zelfmoord(poging) uitvoerden in 2017 werden geselecteerd (n = 574). Een controlegroep met verhoogd risico (n = 207.308) werd geselecteerd uit patiënten met psychische problematiek maar zonder een zelfmoordpoging in 2017. De dataset werd opgesplitst in een trainingset en testset om een AI-model dat zelfmoordneigingen tracht te detecteren een risicostratificatie uitvoert, te ontwikkelen en te testen.	65% van de casussen uit de steekproef kwam op consultatie bij hun huisarts in de 30 dagen voor de zelfmoord(poging). Gebruik van het AI-model gaf een positieve voorspellende waarde van 0,05 (BI 0,04–0,06). Daarnaast gaf het een sensitiviteit van 0,39 (BI 0,32–0,47) en een AUC van 0,85 (BI 0,81–0,88). Bijna alle patiënten uit de controlegroep die risico liep werden correct geclassificeerd (specificiteit = 0,98 (0,97–0,98)). Uit 20 patiënten uit de groep die risico liep en door het AI-model als hoog risico beschouwd werd, pleegden twee een zelfmoord(poging).
Wang et al. (48) O en V ML Diagnostiek Spirometrie Pneumologie	Voor de ontwikkeling van het model werd gebruikgemaakt van spirometrie-PDF-bestanden die tussen oktober 2017 en oktober 2020 waren verkregen uit één ziekenhuis. Er werd gebruikgemaakt van *ATS*/*ERS* 2019-criteria om de kwaliteit van de meting vast te stellen en eventuele fouten te identificeren. Vervolgens werden de metingen verdeeld in trainings- en interne testsets. In totaal werden 15.693 bestanden gelabeld, waarvan 4.592 bestanden werden toegewezen aan de interne testset en de rest de trainingsset vormde. Extra bestanden (182) van drie andere ziekenhuizen werden gebruikt als externe testset. Na de ontwikkeling van de toepassing om de kwaliteit van en fouten in de metingen te detecteren, werd deze gevalideerd bij 30 huisartsen. De studie duurde 3 maanden. Huisartsen, opgeleid om spirometrie uit te voeren, werden geëvalueerd. Er waren drie beoordelingsmomenten, één bij de start van de studie (maand 0) en dan telkens na 1 maand. (maand 1 en maand 2). De huisartsen kregen pas toegang tot de toepassing vanaf maand 1. Er werd een onderscheid gemaakt tussen acceptabele, bruikbare (inclusief acceptabele) en goede curves.	De maandelijkse percentages acceptabele en bruikbare metingen waren significant hoger in maand 1 en maand 2 dan op maand 0. In maand 2, met de AI-ondersteuning, werden gemiddelden voor acceptabele metingen gezien van 91,8% voor FEV1 en 89,4% voor FVC. In tegenstelling hiermee vertoonde de maand 0 lagere gemiddelde percentages (respectievelijk 81,9% en 70,2%) (beide p < 0,0001). Een gelijkaardige verbetering werd gezien in de percentages bruikbare curves voor FEV1- en FVC (88,3% en 88,9% in maand 0, 99,2% en 99,2% in maand 2) (beide p < 0,0001). De hogere percentages van metingen met goede kwaliteit voor FEV1 en FVC in maand 1 (respectievelijk 79,1% en 72,3%) en maand 2 (91,8% en 89,0%) tonen een grote verbetering in vergelijking met maand 0 (respectievelijk 70,8% en 52,8%) (beide P < 0,0001).

Yao et al. (49) V ML Doorverwijzing en diagnostiek Hartfalen (lage ejectiefractie <50%) Cardiologie	De dataset bestond uit ECG's van 22.641 volwassenen (11.068 in de controlegroep en 11.573 in de interventiegroep, de groep met toegang tot een AI-model) afkomstig uit de praktijk van de deelnemende artsen. Er namen 358 clinici deel aan de studie, waarvan 65.5% artsen, 12.4% assistenten en 22.1% verpleegkundigen. Van de artsen waren er 258 huisartsen en 98 internisten. Er werden twee groepen opgesteld, een groep die toegang had tot de resultaten van de AI-toepassing (interventiegroep) en een groep zonder toegang (controlegroep). De toepassing werd wel gebruikt bij beide groepen, maar de controlegroep had geen toegang tot de resultaten. Er werd gekeken naar het verschil in verwijzing voor echocardiografie en bijhorende diagnostiek van lage ejectiefractie (EF).	In de volledige cohort verhoogde de interventie het aantal diagnoses van een lage EF van 1,6% in de controlegroep naar 2,1% in de interventiegroep (OR 1,32 (1,08–1,61), p = 0,007). Bij de 1.356 patiënten met een positief resultaat verbeterde het gebruik van het AI model de diagnose van een lage EF van 14,5% in de controlegroep naar 19,5% in de interventiegroep (OR 1,43 (1,08–1,91), p = 0,01). Er was slechts één significante interactie: de interventie verhoogde de diagnose van een lage EF in sterkere mate bij degenen bij wie de ECG's werden uitgevoerd in poliklinieken. Voor patiënten bij wie een ECG werd uitgevoerd in de poliklinische setting, verhoogde de interventie de diagnose van een lage EF (1,0% voor controle versus 1,6% voor interventie, OR 1,71 (1,23–2,37), p = 0,001).
Yu et al. (74) O en V ML Diagnostiek Psoriasis capitis en seborroïsche keratose Dermatologie	1358 dermatoscopische afbeeldingen van 617 patiënten met gediagnosticeerde huidaandoeningen (psoriasis en seborroïsch eczeem) vormden de dataset. Ze werden willekeurig toegewezen aan de training-, validatie- en testset (1,088/134/136). De diagnostische accuraatheid van een AI-gebaseerd model werd vergeleken met de diagnostische accuraatheid van zes dermatologen en twee huisartsen.	De behaalde accuraatheid, in termen van de AUC, was beter voor het AI-model dan voor de verschillende dermatologen. Als de voorspellingen van het AI-model beschikbaar werden gesteld, steeg de accuraatheid van één dermatoloog in opleiding en twee huisartsen significant: de AUC steeg van 0.600, 0.537 en 0.575 naar respectievelijk 0.849, 0.778 en 0.778. Kappawaarden stegen van 0.191, 0.071, en 0.143 naar respectievelijk 0.679, 0.550 en 0.568.
Zafari et al. (101) O NLP en ML Diagnostiek Posttraumatisch stresssyndroom (PTSS) Psychiatrie	De dataset bestond uit gegevens van huisartsen die lid waren van het *Manitoba Primary Care Research Network.* Het betreft 266 huisartsen uit 48 centra in Manitoba, Canada. Patiëntengegevens van 154.118 patiënten waren beschikbaar. De gegevens bevatten zaken zoals demografische informatie, voorschriften, ICD-9 codes alsook notities. Het laatste wordt beschouwd als ongestructureerde data, terwijl de eerste gestructureerd zijn. Een steekproef van deze gegevens van 195 patiënten bij wie de PTSS-diagnose was bevestigd en van de patiënten zonder PTSS, werden geïncludeerd. Deze werden opgesplitst in een training-, validatie- en testset. Er werden twee modellen	Het sequentiële model leverde de beste resultaten op voor sensitiviteit (0,77), F1-score (0,76) en AUC (0,88). Het parallelle model behaalde de hoogste positieve voorspellende waarde (0,75).

	ontwikkeld die de gestructureerde en ongestructureerde data enerzijds sequentieel en anderzijds parallel verwerkten.	
Zhang et al. (75) O en V ML Diagnostiek Vitiligo Dermatologie	Een interne dataset (n=2,876) en een publieke dataset (n=1,341) met afbeeldingen van vitiligo en andere gepigmenteerde/gehypopigmenteerde laesies werden gebruikt. De interne dataset bestond uit afbeeldingen, retrospectief verkregen, van patiënten die de dermatologieafdeling van het *Qingdao Women and Children's Hospital (QWCH)* in China bezochten. De openbare dataset werd samengesteld uit 7 openbare dermatologie-datasets. Beide datasets werden opgesplitst in een training- en validatieset en testset. Drie modellen werden getraind op afbeeldingen in beide datasets. De resultaten van de modellen werden vergeleken met die van 14 artsen, waaronder 4 dermatologen, 5 dermatologieassistenten en 5 huisartsen. De dermatologen werden opgesplitst volgens hun ervaring. De resultaten van de modellen werden vergeleken met die van de artsen. In tegenstelling tot de AI modellen, had deze groep van 14 personen ook toegang tot afbeeldingen van de huisdlaesies belicht met de *Wood's lamp*, indien die beschikbaar waren.	Op de testsets met interne data behaalde het beste AI model een F1 score die deze van de meer ervaren dermatologen benaderde, met op de twee testsets respectievelijk een score met standaardafwijking van 0,8864 [0,005] en 0.7649 [0.015] versus 0,8933 [0,044] en 0.7708 [0.085]. Het scoorde beter dan de minder ervaren dermatologen (0,7603 [0,029] en 0.7193 [0.021]), dermatologie-assistenten (0,6161 [0,068] en 0.6243 [0.074]) en huisartsen (0,4964 [0,139] en 0.4633 [0.191[). Op de testset van de publieke dataset waren de AI modellen superieur ten opzichte van de meer ervaren dermatologen (0.9684 [0.005] vs 0.9221 [0.031])

Discussie

De toepassing van AI binnen de huisartsgeneeskunde: de huidige stand van zaken

De huidige studie beoogde om een algemeen overzicht te geven met betrekking tot het eerdere onderzoek over AI-toepassingen in de huisartsenpraktijk. Aan de hand van een scoping review werd de volgende onderzoeksvraag onderzocht: "Welke toepassingen van AI in de huisartsenpraktijk zijn er onderzocht en wat zijn de resultaten?".

Na een systematisch selectieproces werden 68 studies, gepubliceerd tussen januari 2018 en november 2023, geïncludeerd. Het aantal publicaties steeg exponentieel tussen 2018 en 2023. 41% van de studies had als doel om AI-toepassingen te ontwikkelen. 24% van de studies beoogde om de AI-toepassingen te valideren, en 35% van de studies beoogde een combinatie van beide. Hoewel ongeveer de helft van de studies dus als doel had om reeds ontwikkelde AI-modellen te valideren in een nieuwe dataset, moet er opgemerkt worden dat AI-toepassingen slechts in tien studies (15%) prospectief werden onderzocht. Deze resultaten bevestigen de resultaten van Abbasgholizadeh Rahimi et al. (1), een systematische scoping review die studies gepubliceerd voor februari 2020 includeerde. Uit de huidige scoping review blijkt dat zes van de tien prospectieve studies werden uitgevoerd in 2022-2023. Deze trend is positief, aangezien prospectieve studies noodzakelijk zijn alvorens een AI-toepassing effectief kan geïmplementeerd worden in de praktijk.

Verder toont deze studie aan dat de meerderheid (72%) van de AI-toepassingen beoogde om diagnostiek te ondersteunen. Daarnaast vonden de AI-modellen vooral hun toepassing binnen cardiologie (21%), dermatologie (16%) en geriatrie (10%). Tot slot werd een grote variabiliteit opgemerkt in de manier waarop studies methodologie en resultaten rapporteren.

Kritische reflectie en implicaties voor toekomstig onderzoek

Veel van de studies leverden interessante en nuttige resultaten op. Toch kunnen, als een kritische blik geworpen wordt op de methodologie en resultaten van de studies, een aantal veel voorkomende tekortkomingen opgelijst worden. In wat volgt worden deze verder uitgediept.

A. Steekproef

Een groot deel van de studies gaat over het ontwikkelen van AI-toepassingen. Deze studies extraheren vaak gegevens uit databanken en/of ziekenhuisafdelingen, om het AI-model op te trainen en testen. Deze dataset is echter niet altijd representatief voor de huisartsenpraktijken, zoals bijvoorbeeld de dataset gebruikt in de studies van Yu et al. (74), Nemlander et al. (85), en Hung et al. (93). Verder zijn de populaties uit de datasets niet altijd representatief voor andere demografische groepen. Zo gebruikt de studie van Chen et al. (51) gegevens verzameld in Pudong-district van Shanghai. Deze populatie is mogelijk niet representatief voor de algemene bevolking. Tang et al. (29) benadrukken echter in hun recente review het belang van het betrekken van verschillende populaties, zowel bij het trainen als bij het testen van een AI-model, gezien de onvoorspelbaarheid van

demografische variabelen en klinische gegevens op de diagnostische accuraatheid van AI-modellen.

B. Belang van training-, validatie- en testset

De training- en validatie- en testset zijn van cruciaal belang bij de ontwikkeling van een AI-model. De studie van Abascal Azanza et al. (36) gebruikte bijvoorbeeld geen validatieset bij de training van het AI-model. Dit houdt in dat ze de keuze voor een bepaald algoritme maken op basis van de testset en vervolgens ook de scores op die testset gebruiken als hun eindresultaat. De vraag blijft of het model dan niet eerder toevallig op dat deel van de data goed werkt. Het gebruik van een aparte testset, die niet gebruikt wordt om het uiteindelijk algoritme te selecteren, staat toe om een eerlijkere score te geven.

Ook zien we dat de resultaten van de toepassing beter zijn indien de testset vergelijkbaar is met de trainingsset. In de studies Muñoz-López et al. (107) en Escalé-Besa et al. (43) bijvoorbeeld, zien we dat de resultaten van het model slechter waren op een nieuwe dataset. Indien men dit echter corrigeerde voor de diagnoses waarop het model expliciet getraind was, verhoogde de prestatie.

Als de prevalentie van bepaalde aandoeningen in de trainingset te laag is, dan behaalt het AI-model in een testset voor de diagnose van die specifieke aandoeningen typisch slechtere resultaten. In de studie van Livingstone et al. (106) wordt bijvoorbeeld aangetoond dat het AI-model slecht presteerde voor de identificatie van zeldzame huidaandoeningen vanwege de lage prevalentie in de trainingset. Ook de studie van Cha et al. (102) toont het belang hiervan aan. Deze onderzoekers ontwikkelden twee AI-modellen, en onderzochten deze in een gebalanceerde (= gelijk aantal beelden per diagnose) en niet-gebalanceerde dataset (beelden per diagnose afhankelijk van prevalentie). De diagnostische accuraatheid van deze modellen was variabel naargelang data al dan niet gebalanceerd was in de testset, in tegenstelling tot de diagnostische accuraatheid van artsen. Ongebalanceerde data komt wel vaker voor in datasets, waaronder ook de *HAM10000-dataset*, die in dermatologie-toepassingen gebruikt wordt, zoals in de studie van Lucius et al. (71). In deze dataset vormen foto's van melanomen de meerderheid van de foto's van gevaarlijke huidletsels. Hierdoor is het mogelijk dat er een neiging is om sneller huidletsels te bestempelen als melanoom.

Soms bleek de data ook reeds op voorhand geselecteerd om het model een betere prestatie te geven, zoals bij Malhotra et al. (84) waarbij maligniteiten met overlappende symptomen/risicofactoren aan de diagnose pancreascarcinoom, werden uitgesloten. Een ander voorbeeld is de studie van Ristanoski et al. (87), waar ze probeerden met hun AI-model longkanker te voorspellen op basis van laboresultaten. Ze betrokken echter enkel patiënten (met of zonder achteraf vastgestelde longkanker) die laboresultaten hadden waarvan de waarde buiten het normale bereik viel. Dit impliceert dat de patiënten met longkanker zonder afwijkende laboresultaten per definitie gemist worden.

Tot slot is de norm, "het juiste antwoord", waarmee de AI-modellen getraind worden, van belang. Zo gebruiken sommige studies, bijvoorbeeld de studie van Jammeh et al (79), *read codes* (ICPC-codes) als norm om het AI-model te trainen. Dit is echter nog onvoldoende, aangezien ongecodeerde gegevens worden gemist en bijgevolg niet alle mogelijke variabelen worden benut. Een ander voorbeeld is de studie van Dormosh et al. (56), waarin een AI-model werd ontwikkeld om vallen bij geriatrische patiënten te voorspellen. Hiervoor werden echter enkel vallen met medisch consultatie geïncludeerd.

Onderzoek (109) toont echter aan dat vallen vaak niet gedetecteerd wordt door artsen. Wanneer het model zou worden getraind op meer volledige data, bijvoorbeeld door een studie waarbij vallen prospectief geregistreerd wordt, zou de meerwaarde voor de klinische praktijk groter zijn.

C. Toepasbaarheid en klinische relevantie

De onderzochte AI-toepassingen zijn niet altijd klinisch relevant. De studie van Ozturk et al. (52) is hier een voorbeeld van. In deze studie werd een toepassing ontwikkeld dat het risico op hypertensie bij Type-2 diabetespatiënten evalueert. Bij diabetes is bloeddrukcontrole essentieel. Er lijkt geen klinische meerwaarde om deze groep patiënten nog te verfijnen op hoger risico, want bloeddruk moet sowieso opgevolgd worden in deze patiëntenpopulatie. Het is van belang dat de studieopzet van toekomstig onderzoek multidisciplinair wordt ontwikkeld. Een samenwerking tussen experten in AI, zoals ingenieurs, en artsen is belangrijk om een klinisch relevant doel van de studie te formuleren.

Daarnaast kunnen we de bedenking maken dat het gebruik van een AI-toepassing er mogelijks voor kan zorgen dat artsen minder differentieel diagnostisch gaan redeneren. In de studie van Herter et al. (45) werd een AI-toepassing geïmplementeerd en onderzocht, die voor urineweginfecties advies gaf aan artsen met betrekking tot antibioticakeuze. Hoewel deze studie de diagnostische meerwaarde van deze AI-toepassing aantoont, zou het gebruik hiervan er eventueel voor kunnen zorgen dat artsen te veel zouden vertrouwen op deze toepassing, en differentiaal diagnoses uit het oog verliezen waarvoor het AI-model niet getraind is zoals seksueel overdraagbare aandoeningen in dit geval. Het klinisch redeneren blijft belangrijk.

Verder richten validatieonderzoeken zich meestal op het evalueren van de accuraatheid van AI-toepassingen. Daarnaast moet er meer onderzoek gebeuren, dat ook de *feasibility* of toepasbaarheid van de toepassing in de praktijk nagaat, zoals bijvoorbeeld in de studie van Gregoor et al. (110). In deze studie werd niet enkel de accuraatheid van de AI-toepassing in kaart gebracht, maar ook in welke mate huisartsen deze toepassing effectief gebruikten en of ze hun hypothese of advies veranderden na het gebruik van de toepassing. Een ander voorbeeld is de studie van Jain et al. (68) waar naast diagnostische accuraatheid, ook bijvoorbeeld de duur van consultatie en gevoel van vertrouwen in diagnostiek , werden in kaart gebracht.

Tot slot ontbreekt een nadere inspectie van de interne werking van het model in sommige studies, hoewel dit klinische inzichten kan geven aan artsen. Surenda et al. (61) gebruiken bijvoorbeeld een Random Forest algoritme om hartfalen te voorspellen op basis van ECG data. Dit soort algoritme maakt een voorspelling met een aantal parameters. Hoewel het mogelijk is om het belang van de verschillende parameters in de uiteindelijke voorspelling te evalueren, doen ze dit niet in hun studie. Dit is echter wel klinisch relevante informatie. Kennis van de manier waarop een AI-model tot een beslissing komt, zou, naast het leveren van klinische informatie, ook vertrouwen kunnen geven aan artsen om op een veilige manier de AI-modellen te kunnen gebruiken.

D. Onvoldoende diagnostische accuraatheid

Er zijn een aantal studies over de ontwikkeling van AI-toepassingen, waarbij de voorspellende waarde onvoldoende is, zoals bijvoorbeeld in de studie van Bennis et al. (39),

Kanaginsam et al. (91) en Kennedy et al. (98). In deze studies is het aantal vals positieve casussen te hoog. Dit impliceert dat wanneer men een AI-toepassing gebruikt als screenings- of diagnostisch instrument, er heel wat patiënten ten onrechte geïdentificeerd zouden worden. Dit zorgt voor meer medische onderzoeken en kosten voor de maatschappij en ook onnodige bezorgdheid bij de betrokken patiënten. Zoals met elk instrument zal het noodzakelijk zijn een geschikte balans te vinden. Ondanks de lage accuraatheid of positief voorspellende waarde, kan de toepassing toch een meerwaarde bieden. Zo vonden Dros et al. (103) een zeer lage positief voorspellende waarde maar zeer goede negatief voorspellende waarde. In de context van zeldzame ziekten heeft deze toepassing eventueel een meerwaarde om de diagnose uit te sluiten. Hoe de resultaten van specifieke AI-toepassingen kunnen geïnterpreteerd worden, moet dan ook uitgebreid uitgelegd worden aan zorgverleners die deze toepassingen zouden gebruiken in de huisartsenpraktijk.

Verder rapporteert een groot deel van deze studies een hoge AUC binnen de testset. De studies van Escalé-Besa et al. (43), Kennedy et al. (98) en Muñoz-López et al. (107) tonen echter aan dat de resultaten van AI-modellen sterk kunnen dalen in opvolgstudies, wanneer het getrainde AI-model wordt getest in een nieuwe onafhankelijke dataset. Dit in tegenstelling tot het AI-model in de studie van Parente et al. (89), dat gelijkaardige resultaten opleverde wanneer het reeds ontwikkelde model (op basis van data uit spoedafdelingen) werd toegepast op data uit de eerstelijnszorg.

Ook is het belangrijk om de accuraatheid te beoordelen per diagnose. Zo vonden Tschandl et al. (70) dat indien men keek naar de gemiddelde scores, rekening houdend met benigne en maligne letsels, de AI-toepassing beter scoorde dan artsen. Indien men enkel keer naar de maligne letsels, scoorden de artsen wel beter. Bij AI-toepassingen die differentieel diagnostisch werken, is het dus essentieel om niet enkel een algemeen goede prestatie te hebben, maar ook per diagnose.

E. Vergelijking prestatie

Een minderheid van de studies vergelijkt de resultaten van AI-toepassingen met de resultaten van artsen. Dit is echter wel nodig om de meerwaarde van de AI-toepassing in kaart te brengen.

Sommige geïncludeerde studies doen dit wel, maar de gebruikte methode om de prestatie van artsen te bepalen is niet altijd hoe het momenteel gebeurt in de huisartspraktijk. In de studie van Hart et al. (86) werden de resultaten van het AI-model bijvoorbeeld vergeleken met de resultaten van 15 artsen. De patiënten uit deze studie vulden een vragenlijst in, en op basis van hun antwoorden classificeerden de artsen de patiënten in twee groepen. Een groep met een verhoogd risico op endometriumcarcinoom en een groep met een verlaagd risico. In de huisartspraktijk bepalen we echter niet of een patiënt een verhoogd risico heeft op basis van één vragenlijst, maar ook op basis van ervaring, kennis en andere beschikbare gegevens. Toekomstig onderzoek moet bepalen of dit model effectief beter presteert dan de huidige norm in de huisartspraktijk, zoals bijvoorbeeld in de studie van Herter et al. (45), waarbij een AI-tool werd geïmplementeerd in huisartspraktijken, en outcome (het succesvol behandelen van urineweginfecties) werd vergeleken met andere huisartsenpraktijken die deze tool niet gebruikten.

Verder valt het op dat de meerderheid van de studies die resultaten van AI-toepassingen vergelijken met de resultaten van artsen, slechts beroep doen op een

beperkt aantal artsen in deze laatste groep. Bijvoorbeeld Hirosawa et al.(104) en Cha et al.(102)

Tot slot richten te weinig studies zich op de vergelijking in resultaten tussen huisartsen, de AI-toepassing en de combinatie van beide. Dit soort onderzoek zou meer zicht geven op de meerwaarde van de AI-toepassing in de praktijk. In het onderzoek van Himmelich et al.(62) worden bijvoorbeeld de resultaten van de AI-toepassing wel vergeleken met de resultaten van de artsen, maar het is ook nodig om de resultaten van de toepassing te vergelijken bij huisartsen met AI en zonder AI. Dit werd wel gedaan bij Lucius et al. (71) en Anderson et al.(96), waarbij de artsen in tweede tijd toegang kregen tot de AI-toepassing.

F. Patiënt als gebruiker

De grote meerderheid van de geïncludeerde studies handelt over AI-toepassingen te gebruiken door huisartsen. Oude Nijeweme-d'Hollosy et al. (50) beoogden echter om een AI-toepassing te ontwikkelen, die patiënten met lage rugpijn zelf kunnen gebruiken om advies te krijgen rond doorverwijzing. Er zijn echter gevaren die we moeten benadrukken bij het gebruik van AI door patiënten voor medische doeleinden. Uit de studie van Gregoor et al. (110) blijkt bijvoorbeeld dat 54% van de patiënten voor wie de app aangaf dat een huidletsel met hoge waarschijnlijkheid goedaardig was hun afspraak bij de huisarts zouden afzeggen met dit resultaat. Vals negatieve resultaten zouden in zo'n casussen dus niet gezien zijn door een huisarts, die naast het gebruik van de AI-toepassing ook zijn kennis en ervaring zou gebruiken voor doorverwijzing.

G. Retrospectieve studies

De meerderheid van validatiestudies onderzocht casussen retrospectief zoals bijvoorbeeld de studie van Jain et al. (68) waarbij huisartsen retrospectief verschillende casussen met betrekking tot huidaandoeningen beoordelen. Hoewel dit een eerste veelbelovende stap is, zijn er enkele tekorten die gerelateerd zijn aan dit retrospectief onderzoeksdesign. Zo krijgen de artsen bijvoorbeeld al voorgeselecteerde informatie, wat niet te vergelijken is met een consultatie bij de huisarts. Meer prospectieve studies zijn nodig, om de resultaten uit eerder onderzoek met retrospectieve data te repliceren.

Limitaties

De huidige scoping review kent echter ook enkele limitaties. Hoewel het aangeraden is om de studieselectie door twee onafhankelijke beoordelaars te laten gebeuren (40), was dit niet mogelijk in het kader van deze masterproef. De studies in de huidige scoping review werden dus geselecteerd door één beoordelaar. De methodologische systematische review van Waffenschmidt et al. (111) beschrijft echter dat er hierdoor mogelijks studies gemist zijn.

Verder werd de kwaliteit van de geïncludeerde studies niet op een systematische manier beoordeeld. Hoewel dit niet noodzakelijk is voor het uitvoeren van een scoping review (40), kan dit wel relevante informatie opleveren.

Tot slot werden alle kwalitatieve en kwantitatieve studies geïncludeerd in deze scoping review, met de bedoeling om eerdere literatuur zo breed mogelijk in kaart te brengen. Ook *conference papers* die preliminaire resultaten rapporteren, zoals bijvoorbeeld de studie van Judd et al. (94), werden geïncludeerd. Deze resultaten moeten met de nodige voorzichtigheid geïnterpreteerd worden.

Conclusie

Na een systematisch selectieproces werden 68 studies geïncludeerd om meer zicht te krijgen op de huidige stand van zaken met betrekking tot AI-toepassingen in de huisartsgeneeskunde. Het aantal publicaties steeg exponentieel tussen 2018 en 2023. 41% van de geïncludeerde studies had als doel om AI-toepassingen te ontwikkelen. 24% van de studies beoogde om de AI-toepassingen te valideren, en 35% van de studies deed een combinatie van beide. Er waren slechts tien studies (15%) die AI-toepassingen prospectief onderzochten. Verder beoogde de meerderheid (72%) van de AI-toepassingen om diagnostiek te ondersteunen. Daarnaast kende de meerderheid van de AI-modellen hun toepassing binnen cardiologie (21%), dermatologie (16%) en geriatrie (10%). Een grote variabiliteit werd opgemerkt in de manier waarop methodologie en resultaten gerapporteerd werden.

Hoewel de meeste geïncludeerde studies veelbelovende toepassingen van AI onderzochten, die het werk van huisartsen mogelijks efficiënter en of accurater kunnen maken, zijn er nog meer stappen nodig om tot een effectieve implementatie te komen. Bovenstaande reflectie maakt duidelijk dat het huidige onderzoek naar AI-toepassingen nog in zijn kinderschoenen staat, en dat meer prospectief validatie onderzoek nodig is om de toepasbaarheid, accuraatheid en meerwaarde voor deze toepassingen te onderzoeken, alvorens ze kunnen gebruikt worden in de praktijk. Uit de huidige scoping review blijkt dat zes van de tien prospectieve studies werden uitgevoerd in 2022-2023. Dit toont een positieve trend aan in het onderzoeksveld.

Referenties

1. Abbasgholizadeh Rahimi S, Légaré F, Sharma G, Archambault P, Zomahoun HTV, Chandavong S, e.a. Application of artificial intelligence in community-based primary health care: systematic scoping review and critical appraisal. J Med Internet Res. 2021;23(9):e29839.
2. Russell SJ, Norvig P. Artificial intelligence a modern approach. London; 2010.
3. Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press; 2016.
4. Bradshaw TJ, Huemann Z, Hu J, Rahmim A. A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging. Radiol Artif Intell. 2023;e220232.
5. Schmidhuber J. Annotated history of modern AI and Deep learning. ArXiv Prepr ArXiv221211279. 2022;
6. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys. december 1943;5(4):115-33.
7. Widrow B. Adaptive" adaline" Neuron Using Chemical" memistors.". 1960.
8. Shortliffe EH. Mycin: A knowledge-based computer program applied to infectious diseases. In: Proceedings of the Annual Symposium on Computer Application in Medical Care. American Medical Informatics Association; 1977. p. 66.
9. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012;25.
10. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, e.a. Mastering the game of Go with deep neural networks and tree search. nature. 2016;529(7587):484-9.
11. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv Prepr ArXiv181004805. 2018;
12. Reddy MDM, Basha MSM, Hari MMC, Penchalaiah MN. Dall-e: Creating images from text. UGC Care Group J. 2021;8(14):71-5.
13. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. 2018;
14. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI Blog. 2019;1(8):9.
15. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, e.a. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877-901.
16. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, e.a. Generative adversarial nets. Adv Neural Inf Process Syst. 2014;27.
17. Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R. Nerf: Representing scenes as neural radiance fields for view synthesis. Commun ACM. 2021;65(1):99-106.
18. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Adv Neural Inf Process Syst. 2020;33:6840-51.
19. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, e.a. Attention is all you need. Adv Neural Inf Process Syst. 2017;30.
20. Tu T, Azizi S, Driess D, Schaekermann M, Amin M, Chang PC, e.a. Towards generalist biomedical AI. ArXiv Prepr ArXiv230714334. 2023;
21. ChatGPT. ChatGPT Release Notes [Internet]. 2023. Beschikbaar op: https://help.openai.com/en/articles/6825453-chatgpt-release-notes#h_4799933861
22. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. J Am Med Inform Assoc JAMIA. 1 oktober 2018;25(10):1419-28.
23. Nathan Benaick, Air Street Capital Team. State of AI Report 2023 [Internet]. 2023 dec.

Beschikbaar op: https://www.stateof.ai/
24. Dovey SM. A preliminary taxonomy of medical errors in family practice. Qual Saf Health Care. 1 september 2002;11(3):233-8.
25. Beard JR, Bloom DE. Towards a comprehensive public health response to population ageing. The Lancet. februari 2015;385(9968):658-61.
26. Lin SY, Mahoney MR, Sinsky CA. Ten Ways Artificial Intelligence Will Transform Primary Care. J Gen Intern Med. augustus 2019;34(8):1626-30.
27. Hazarika I. Artificial intelligence: opportunities and implications for the health workforce. Int Health. 1 juli 2020;12(4):241-5.
28. Willmen T, Völkel L, Ronicke S, Hirsch MC, Kaufeld J, Rychlik RP, e.a. Health economic benefits through the use of diagnostic support systems and expert knowledge. BMC Health Serv Res. december 2021;21(1):947.
29. Tang L, Li J, Fantus S. Medical artificial intelligence ethics: A systematic review of empirical studies. Digit Health. januari 2023;9:20552076231186064.
30. Wewetzer L, Held LA, Goetz K, Steinhäuser J. Determinants of the implementation of artificial intelligence-based screening for diabetic retinopathy—a cross-sectional study with general practitioners in Germany. Digit Health. januari 2023;9:205520762311766.
31. Glock H, Milos Nymberg V, Borgström Bolmsjö B, Holm J, Calling S, Wolff M, e.a. Attitudes, barriers, and concerns regarding telemedicine among Swedish primary care physicians: a qualitative study. Int J Gen Med. 2021;9237-46.
32. Dumić-Čule I, Orešković T, Brkljačić B, Kujundžić Tiljak M, Orešković S. The importance of introducing artificial intelligence to the medical curriculum–assessing practitioners' perspectives. Croat Med J. 2020;61(5):457-64.
33. Buck C, Doctor E, Hennrich J, Jöhnk J, Eymann T. General Practitioners' Attitudes Toward Artificial Intelligence–Enabled Systems: Interview Study. J Med Internet Res. 27 januari 2022;24(1):e28916.
34. Blease C, Kaptchuk TJ, Bernstein MH, Mandl KD, Halamka JD, DesRoches CM. Artificial intelligence and the future of primary care: exploratory qualitative study of UK general practitioners' views. J Med Internet Res. 2019;21(3):e12802.
35. Hogarty DT, Su JC, Phan K, Attia M, Hossny M, Nahavandi S, e.a. Artificial Intelligence in Dermatology—Where We Are and the Way to the Future: A Review. Am J Clin Dermatol. februari 2020;21(1):41-7.
36. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, e.a. Artificial Intelligence in Cardiology. J Am Coll Cardiol. juni 2018;71(23):2668-79.
37. Ueda D, Shimazaki A, Miki Y. Technical and clinical overview of deep learning in radiology. Jpn J Radiol. januari 2019;37(1):15-33.
38. Kueper JK, Terry AL, Zwarenstein M, Lizotte DJ. Artificial Intelligence and Primary Care Research: A Scoping Review. Ann Fam Med. mei 2020;18(3):250-8.
39. Turcian D, Stoicu-Tivadar V. Artificial intelligence in primary care: an overview. Stud Health Technol Inf. 2022;289:208-11.
40. Kazi M, Chowdhury N, Chowdhury M, Turin T. Conducting comprehensive scoping reviews to systematically capture the landscape of a subject matter. Popul Med. 23 december 2021;3(December):1-9.
41. Calcagno S, Biondizoccai G, Stankovic T, Szabo E, Szabo AB, Kecskes I. Novel tech throws knock-out punch to ECG improving GP referral decisions to cardiology. Open Heart [Internet]. 2022;9(1). Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85125741140&doi=10.1136%2fopenhrt-2021-001852&partnerID=40&md5=2b4ee3d359389b941f05bf263cc9aa2b
42. Cruz NP, Canales L, Muñoz JG, Pérez B, Arnott I. Improving adherence to clinical pathways through natural language processing on electronic medical records. In 2019. p. 561-5. Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85071449503&doi=10.3233%2fSHTI190285&partnerID=40&md5=ab069675c16ea37bc530439e5221f1f2
43. Escalé-Besa A, Yélamos O, Vidal-Alaball J, Fuster-Casanovas A, Catalina Q, Börve A,

e.a. Exploring the potential of artificial intelligence in improving skin lesion diagnosis in primary care. Sci Rep. 2023;13(1).
44. Gregoor A, Sangers T, Eekhof J, Howe S, Revelman J, Litjens R, e.a. Artificial intelligence in mobile health for skin cancer diagnostics at home (AIM HIGH): a pilot feasibility study. ECLINICALMEDICINE. 2023;60.
45. Herter W, Khuc J, Cinà G, Knottnerus B, Numans M, Wiewel M, e.a. Impact of a Machine Lear ning-Based Decision Support System for Urinary Tract Infections: Prospective Observational Study in 36 Primary Care Practices. JMIR Med Inform. 2022;10(5).
46. Hill NR, Arden C, Beresford-Hulme L, Camm AJ, Clifton D, Davies DW, e.a. Identification of undiagnosed atrial fibrillation patients using a machine learning risk prediction algorithm and diagnostic testing (PULsE-AI): Study protocol for a randomised controlled trial. Contemp Clin Trials [Internet]. 2020;99. Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85093084431&doi=10.1016%2fj.cct.2020.106191&partnerID=40&md5=b85b21c424d0007c63d336bd509fa746
47. Rushlow D, Croghan I, Inselman J, Thacher T, Friedman P, Yao X, e.a. Clinician Adoption of an Artificial Intelligence Algorithm to Detect Left Ventricular Systolic Dysfunction in Primary Care. MAYO Clin Proc. 2022;97(11):2076-85.
48. Wang Y, Li Y, Chen W, Zhang C, Liang L, Huang R, e.a. Deep learning for spirometry quality assurance with spirometric indices and curves. Respir Res. 2022;23(1).
49. Yao X, Rushlow D, Inselman J, McCoy R, Thacher T, Behnken E, e.a. Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. Nat Med. 2021;27(5):815-+.
50. Oude Nijeweme-d'Hollosy W, van Velsen L, Poel M, Groothuis-Oudshoorn CGM, Soer R, Hermens H. Evaluation of three machine learning models for self-referral decision support on low back pain in primary care. Int J Med Inf. 2018;110:31-41.
51. Chen N, Fan F, Geng J, Yang Y, Gao Y, Jin H, e.a. Evaluating the risk of hypertension in residents in primary care in Shanghai, China with machine learning algorithms. Front Public Health [Internet]. 2022;10. Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85139973168&doi=10.3389%2ffpubh.2022.984621&partnerID=40&md5=67b88510483bbf85706f3cc5cbc242f5
52. Ozturk B, Lawton T, Smith S, Habli I. Predicting Progression of Type 2 Diabetes Using Primary Care Data with the Help of Machine Learning. In 2023. p. 38-42. Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85159759818&doi=10.3233%2fSHTI230060&partnerID=40&md5=acc3cb30d9e9bb8036b3220f7f0da98b
53. Crutzen S, Belur Nagaraj S, Taxis K, Denig P. Identifying patients at increased risk of hypoglycaemia in primary care: Development of a machine learning-based screening tool. Diabetes Metab Res Rev [Internet]. 2021;37(7). Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85101281245&doi=10.1002%2fdmrr.3426&partnerID=40&md5=ca08c556d8c0f60814864d94ed2f88b3
54. DelPozo-Banos M, John A, Petkov N, Berridge D, Southern K, LLoyd K, e.a. Using Neural Networks with Routine Health Records to Identify Suicide Risk: Feasibility Study. JMIR Ment Health. 2018;5(2).
55. van Mens K, Elzinga E, Nielen M, Lokkerbol J, Poortvliet R, Donker G, e.a. Applying machine learning on health record data from general practitioners to predict suicidality. INTERNET Interv- Appl Inf Technol Ment Behav Health. 2020;21.
56. Dormosh N, Schut M, Heymans M, Maarsingh O, Bouman J, van der Velde N, e.a. Predicting future falls in older people using natural language processing of general practitioners' clinical notes. AGE AGEING. 2023;52(4).
57. Homburg M, Meijer E, Berends M, Kupers T, Hartman T, Muris J, e.a. A Natural Language Processing Model for COVID-19 Detection Based on Dutch General Practice Electronic Health Records by Using Bidirectional Encoder Representations From Transformers: Development and Validation Study. J Med INTERNET Res. 2023;25.

58. Topaz M, Zolnoori M, Norful AA, Perrier A, Kostic Z, George M. Speech recognition can help evaluate shared decision making and predict medication adherence in primary care setting. PLoS ONE [Internet]. 2022;17(8). Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85135422553&doi=10.1371%2fjournal.pone.0271884&partnerID=40&md5=195dfc927f241ebaf1f133ed916c924c
59. Bennis FC, Hoogendoorn M, Aussems C, Korevaar JC. Prediction of heart failure 1 year before diagnosis in general practitioner patients using machine learning algorithms: a retrospective case-control study. BMJ Open [Internet]. 2022;12(8). Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85136960488&doi=10.1136%2fbmjopen-2021-060458&partnerID=40&md5=84fcb002855f915bdb18276a404d9766
60. Guida F, Lenatti M, Keshavjee K, Khatami A, Guergachi A, Paglialonga A. Characterization of Inclination Analysis for Predicting Onset of Heart Failure from Primary Care Electronic Medical Records. Sensors [Internet]. 2023;23(9). Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85159378101&doi=10.3390%2fs23094228&partnerID=40&md5=1ba4f9ef5c90030ebb1f67f5ef90608a
61. Surendra K, Nürnberg S, Bremer JP, Knorr MS, Ückert F, Wenzel JP, e.a. Pragmatic screening for heart failure in the general population using an electrocardiogram-based neural network. ESC Heart Fail. 2023;10(2):975-84.
62. Himmelreich JCL, Harskamp RE. Diagnostic accuracy of the PMcardio smartphone application for artificial intelligence–based interpretation of electrocardiograms in primary care (AMSTELHEART-1). Cardiovasc Digit Health J. 2023;4(3):80-90.
63. Hill NR, Ayoubkhani D, McEwan P, Sugrue DM, Farooqui U, Lister S, e.a. Predicting atrial fibrillation in primary care using machine learning. PLoS ONE [Internet]. 2019;14(11). Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074347946&doi=10.1371%2fjournal.pone.0224582&partnerID=40&md5=a8142e60f6893c6c8ab39d8112cb17aa
64. Nadarajah R, Wu J, Hogg D, Raveendra K, Nakao YM, Nakao K, e.a. Prediction of short-term atrial fibrillation risk using primary care electronic health records. Heart. 2023;109(14):1072-9.
65. Sekelj S, Sandler B, Johnston E, Pollock K, Hill N, Gordon J, e.a. Detecting undiagnosed atrial fibrillation in UK primary care: Validation of a machine learning prediction algorithm in a retrospective cohort study. Eur J Prev Cardiol. 2021;28(6):598-605.
66. Akyea R, Qureshi N, Kai J, Weng S. NPJ Digit Med. 2020;3(1).
67. Krentz AJ, Haddon-Hill G, Zou X, Pankova N, Jaun A. Machine Learning Applied to Cholesterol-Lowering Pharmacotherapy: Proof-of-Concept in High-Risk Patients Treated in Primary Care. Metab Syndr Relat Disord [Internet]. 2023; Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85172130489&doi=10.1089%2fmet.2023.0009&partnerID=40&md5=2acea72c44e2f2886f60807bbef298a4
68. Jain A, Way D, Gupta V, Gao Y, De Oliveira Marinho G, Hartford J, e.a. Development and Assessment of an Artificial Intelligence-Based Tool for Skin Condition Diagnosis by Primary Care Physicians and Nurse Practitioners in Teledermatology Practices. JAMA Netw Open [Internet]. 2021;4(4). Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85105042261&doi=10.1001%2fjamanetworkopen.2021.7249&partnerID=40&md5=79c5774a686c4e31ee9982ed849d8b7e
69. Liu Y, Jain A, Eng C, Way D, Lee K, Bui P, e.a. A deep learning system for differential diagnosis of skin diseases. Nat Med. 2020;26(6):900-+.
70. Tschandl P, Codella N, Akay B, Argenziano G, Braun R, Cabo H, e.a. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. LANCET Oncol. 2019;20(7):938-47.
71. Lucius M, De All J, De All J, Belvisi M, Radizza L, Lanfranconi M, e.a. Deep Neural Frameworks Improve the Accuracy of General Practitioners in the Classification of

Pigmented Skin Lesions. DIAGNOSTICS. 2020;10(11).
72. Phillips M, Greenhalgh J, Marsden H, Palamaras I. Detection of malignant melanoma using artificial intelligence: An observational study of diagnostic accuracy. Dermatol Pract Concept [Internet]. 2020;10(1). Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85089107599&doi=10.5826%2fdpc.1001a11&partnerID=40&md5=b9e113ba2c101a23c9aaae8c49a92d37
73. Giavina-Bianchi M, de Sousa RM, de Almeida Paciello VZ, Vitor WG, Okita AL, Prôa R, e.a. Implementation of artificial intelligence algorithms for melanoma screening in a primary care setting. PLoS ONE [Internet]. 2021;16(9). Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85115796551&doi=10.1371%2fjournal.pone.0257006&partnerID=40&md5=a1f74f025c015f02cbacf59ce4fe3413
74. Yu Z, Kaizhi S, Jianwen H, Guanyu Y, Yonggang W. A deep learning-based approach toward differentiating scalp psoriasis and seborrheic dermatitis from dermoscopic images. Front Med. 2022;9.
75. Zhang L, Mishra S, Zhang T, Zhang Y, Zhang D, Lv Y, e.a. Design and Assessment of Convolutional Neural Network Based Methods for Vitiligo Diagnosis. Front Med [Internet]. 2021;8. Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85118346579&doi=10.3389%2ffmed.2021.754202&partnerID=40&md5=d5a1f69c0de26df9026526c200a25b74
76. Ford E, Sheppard J, Oliver S, Rooney P, Banerjee S, Cassell J. Automated detection of patients with dementia whose symptoms have been identified in primary care but have no formal diagnosis: a retrospective case-control study using electronic primary care records. BMJ OPEN. 2021;11(1).
77. Ford E, Starlinger J, Rooney P, Oliver S, Banerjee S, van Marwijk H, e.a. Could dementia be detected from UK primary care patients' records by simple automated methods earlier than by the treating physician? A retrospective case-control study. Wellcome Open Res [Internet]. 2020;5. Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85089774887&doi=10.12688%2fwellcomeopenres.15903.1&partnerID=40&md5=b55cebbe70abe0f9cae9e537aba2e414
78. Ford E, Rooney P, Oliver S, Hoile R, Hurley P, Banerjee S, e.a. Identifying undetected dementia in UK primary care patients: a retrospective case-control study comparing machine-learning and standard epidemiological approaches. BMC Med Inform Decis Mak. 2019;19(1).
79. Jammeh EA, Carroll CB, Pearson Stephen W, Escudero J, Anastasiou A, Zhao P, e.a. Machine-learning based identification of undiagnosed dementia in primary care: A feasibility study. BJGP Open [Internet]. 2018;2(2). Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074425970&doi=10.3399%2fbjgpopen18X101589&partnerID=40&md5=d8a5147799472ec69c303406c707517a
80. Maclagan L, Abdalla M, Harris D, Stukel T, Chen B, Candido E, e.a. Can Patients with Dementia Be Identified in Primary Care Electronic Medical Records Using Natural Language Processing? J Healthc Inform Res. 2023;7(1):42-58.
81. Suárez-Araujo CP, García Báez P, Cabrera-León Y, Prochazka A, Rodríguez Espinosa N, Fernández Viadero C. A Real-Time Clinical Decision Support System, for Mild Cognitive Impairment Detection, Based on a Hybrid Neural Architecture. Comput Math Methods Med [Internet]. 2021;2021. Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85110235245&doi=10.1155%2f2021%2f5545297&partnerID=40&md5=970991ddfb7f7857ef1f90cbe9d69955
82. Soerensen P, Christensen H, Laursen S, Hardahl C, Brandslund I, Madsen J. Using artificial intelligence in a primary care setting to identify patients at risk for cancer: a risk prediction model based on routine laboratory tests. Clin Chem Lab Med. 2022;60(12):2005-16.
83. Briggs E, de Kamps M, Hamilton W, Johnson O, McInerney C, Neal R. Machine Learning for Risk Prediction of Oesophago-Gastric Cancer in Primary Care: Comparison with Existing Risk-Assessment Tools. CANCERS. 2022;14(20).

84. Malhotra A, Rachet B, Bonaventure A, Pereira SP, Woods LM. Can we screen for pancreatic cancer? Identifying a sub-population of patients at high risk of subsequent diagnosis using machine learning techniques applied to primary care data. PLoS ONE [Internet]. 2021;16(6). Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85107299743&doi=10.1371%2fjournal.pone.0251876&partnerID=40&md5=4bc3c8703250814c595823dd6f16854d
85. Nemlander E EM Abedi E, Hasselström J, Sjövall A, Carlsson AC, Rosenblad A. A machine learning tool for identifying non-metastatic colorectal cancer in primary care.
86. Hart GR, Yan V, Huang GS, Liang Y, Nartowt BJ, Muhammad W, e.a. Population-Based Screening for Endometrial Cancer: Human vs. Machine Intelligence. Front Artif Intell [Internet]. 2020;3. Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85114785755&doi=10.3389%2ffrai.2020.539879&partnerID=40&md5=825a16f4ca4122f9a477d59e65cbb4f8
87. Ristanoski G, Emery J, Gutierrez JM, McCarthy D, Aickelin U. Primary Care Datasets for Early Lung Cancer Detection: An AI Led Approach. In 2021. p. 83-92. Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85111391901&doi=10.1007%2f978-3-030-77211-6_9&partnerID=40&md5=acc03e96b31ed2c25db43fb2bee35c11
88. Dhanda G, Asham M, Shanks D, O'malley N, Hake J, Satyan MT, e.a. Adaptation and External Validation of Pathogenic Urine Culture Prediction in Primary Care Using Machine Learning. Ann Fam Med. 2023;21(1):11-8.
89. Parente D, Shanks D, Yedlinksy N, Hake J, Dhanda G. Machine Learning Prediction of Urine Cultures in Primary Care. Ann Fam Med [Internet]. 2023;(21). Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85150982819&doi=10.1370%2fafm.21.s1.4141&partnerID=40&md5=93190f77421914350d512ce13d931cf5
90. Pareja-Ríos A, Ceruso S, Romero-Aroca P, Bonaque-González S. A New Deep Learning Algorithm with Activation Mapping for Diabetic Retinopathy: Backtesting after 10 Years of Tele-Ophthalmology. J Clin Med [Internet]. 2022;11(17). Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85137807863&doi=10.3390%2fjcm11174945&partnerID=40&md5=0e64fb8d2b3f77e1cfac0df09a2c9701
91. Kanagasingam Y, Xiao D, Vignarajan J, Preetham A, Tay-Kearney ML, Mehrotra A. Evaluation of Artificial Intelligence-Based Grading of Diabetic Retinopathy in Primary Care. JAMA Netw Open. 2018;1(5):e182665.
92. Abascal Azanza C, Barrio-Barrio J, Ramos Cejudo J, Ybarra Arróspide B, Devoto MH. Development and validation of a convolutional neural network to identify blepharoptosis. Sci Rep [Internet]. 2023;13(1). Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85174281899&doi=10.1038%2fs41598-023-44686-3&partnerID=40&md5=52ee1e184491ad1016f353e1b2419af7
93. Hung JY, Chen KW, Perera C, Chiu HK, Hsu CR, Myung D, e.a. An Outperforming Artificial Intelligence Model to Identify Referable Blepharoptosis for General Practitioners. J Pers Med [Internet]. 2022;12(2). Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85125051873&doi=10.3390%2fjpm12020283&partnerID=40&md5=7b6d4dee7ad283d2ab2d5042989752fb
94. Judd M, Zulkernine F, Wolfrom B, Barber D, Rajaram A. Detecting low back pain from clinical narratives using machine learning approaches. In 2018. p. 126-37. Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85051941499&doi=10.1007%2f978-3-319-99133-7_10&partnerID=40&md5=ab579de918705dcf6d04db0c176e3e46
95. Ellertsson S LH Sigurdsson EL. Artificial intelligence in the GPs office: a retrospective study on diagnostic accuracy.
96. Anderson PG, Baum GL, Keathley N, Sicular S, Venkatesh S, Sharma A, e.a. Deep Learning Assistance Closes the Accuracy Gap in Fracture Detection Across Clinician Types. Clin Orthop. 2023;481(3):580-8.
97. Chung SW, Han SS, Lee JW, Oh KS, Kim NR, Yoon JP, e.a. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta

Orthop. 2018;89(4):468-73.
98. Kennedy J, Kennedy N, Cooksey R, Choy E, Siebert S, Rahman M, e.a. Predicting a diagnosis of ankylosing spondylitis using primary care health records- A machine learning approach. PLoS ONE [Internet]. 2023;18(3). Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85151348182&doi=10.1371%2fjournal.pone.0279076&partnerID=40&md5=5da05bf60a0a1e347f734dce8653a7f8
99. Kocks J, Cao H, Holzhauer B, Kaplan A, FitzGerald J, Kostikas K, e.a. Diagnostic Performance of a Machine Learning Algorithm (Asthma/Chronic Obstructive Pulmonary Disease [COPD] Differentiation Classification) Tool Versus Primary Care Physicians and Pulmonologists in Asthma, COPD, and Asthma/COPD Overlap. J ALLERGY Clin Immunol- Pract. 2023;11(5).
100. Levkovich I, Elyoseph Z. Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physicians. Fam Med Community Health [Internet]. 2023;11(4). Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85174640818&doi=10.1136%2ffmch-2023-002391&partnerID=40&md5=b2f46334e8350a2f968321722a43b296
101. Zafari H, Kosowan L, Zulkernine F, Signer A. Diagnosing post-traumatic stress disorder using electronic medical record data. Health Informatics J. 2021;27(4).
102. Cha D, Pae C, Lee SA, Na G, Hur YK, Lee HY, e.a. Differential Biases and Variabilities of Deep Learning–Based Artificial Intelligence and Human Experts in Clinical Diagnosis: Retrospective Cohort and Survey Study. JMIR Med Inform [Internet]. 2021;9(12). Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85122035831&doi=10.2196%2f33049&partnerID=40&md5=40da04154380007668bc9f1ce4ff2445
103. Dros JT, Bos I, Bennis FC, Wiegersma S, Paget J, Seghieri C, e.a. Detection of primary Sjögren's syndrome in primary care: developing a classification model with the use of routine healthcare data and machine learning. BMC Prim Care [Internet]. 2022;23(1). Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85135691938&doi=10.1186%2fs12875-022-01804-w&partnerID=40&md5=6db20953cf1aa0030dab5cecf67e4d79
104. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study. Int J Environ Res Public Health [Internet]. 2023;20(4). Beschikbaar op: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85148965327&doi=10.3390%2fijerph20043378&partnerID=40&md5=7d532b86e7afa1897ff97c1776ef6c43
105. Lee H, Aslim E, Balamurali B, Ng L, Kuo T, Lin C, e.a. Development and Validation of a Deep Learning System for Sound-based Prediction of Urinary Flow. Eur Urol FOCUS. 2023;9(1):209-15.
106. Livingstone D, Chau J. Otoscopic diagnosis using computer vision: An automated machine learning approach. LARYNGOSCOPE. 2020;130(6):1408-13.
107. Muñoz-López C, Ramírez-Cornejo C, Marchetti M, Han S, Del Barrio-Díaz P, Jaque A, e.a. Performance of a deep neural network in teledermatology: a single-centre prospective diagnostic study. J Eur Acad Dermatol Venereol. 2021;35(2):546-53.
108. Pyne Y, Wong Y, Fang H, Simpson E. Analysis of 'One in a Million' primary care consultation conversations using natural language processing. BMJ Health CARE Inform. 2023;30(1).
109. Rubenstein LZ, Solomon DH, Roth CP, Young RT, Shekelle PG, Chang JT, e.a. Detection and management of falls and instability in vulnerable elders by community physicians. J Am Geriatr Soc. 2004;52(9):1527-31.
110. Gregoor A, Sangers T, Eekhof J, Howe S, Revelman J, Litjens R, e.a. Artificial intelligence in mobile health for skin cancer diagnostics at home (AIM HIGH): a pilot feasibility study. ECLINICALMEDICINE. 2023;60.
111. Waffenschmidt S, Knelangen M, Sieben W, Bühn S, Pieper D. Single screening

versus conventional double screening for study selection in systematic reviews: a methodological systematic review. BMC Med Res Methodol. 2019;19(1):1-9.