De rol van AI in diagnostiek en classificatie binnen de ggz

F.A.F. Jaspers, G.A. van Wingen, K.R. Jongsma, R.M. Bertens, J.S. Legerstee

Achtergrond De toepassing van artificiële intelligentie (AI) biedt nieuwe kansen voor classificatie en diagnostiek in de algemene gezondheidszorg. Deze technologieën stuit in de geestelijke gezondheidszorg echter op extra hindernissen.

Doel De huidige stand van zaken rond AI-toepassingen in psychiatrische diagnostiek verhelderen.

Methode Beschouwend essay waarin wij als specialisten uit psychiatrie, AI-technologie, data-innovatie, ethiek en recht onze inzichten bundelen.

Resultaten De technologie ontwikkelt zich sneller dan de klinische praktijk kan bijbenen. Omdat supervised-machine-learning-modellen en large language models betrouwbare en uniforme labels vereisen – iets wat het huidige classificatiesysteem nog niet consistent biedt – blijft brede klinische toepassing voorlopig beperkt. Het gebruik van AI stelt hoge eisen aan technische, communicatieve en normatieve competenties van zorgprofessionals, en roept belangrijke ethische vragen op. Zolang AI-systemen niet zijn ontwikkeld of goedgekeurd voor medisch gebruik, blijft juridisch gezien grote terughoudendheid geboden.

Conclusie Pas als klinische, ethische én juridische randvoorwaarden zijn vervuld, is implementatie en verdere ontwikkeling in het diagnostische gerechtvaardigd en biedt verdere implementatie reële kansen op snelle, betrouwbare en objectieve diagnostiek in de psychiatrie en daarmee kansen op efficiëntere zorg.

Een recente gerandomiseerde gecontroleerde trial (RCT) in *Nature Medicine* toonde aan dat artsen die gebruikmaakten van een *large language model* (GPT-4) significant beter scoorden op het diagnosticeren van medische casuïstiek op basis van vignetten dan artsen die gebruikmaakten van conventionele bronnen.[1] Dit voedt het idee dat artificiële intelligentie (AI) mogelijk ook kansen biedt voor het classificeren en diagnosticeren binnen de geestelijke gezondheidszorg (ggz). In dit artikel verkennen we de mogelijke toepassing van large language models (LLM's) en *supervised machine learning* binnen de psychiatrische classificatie en diagnostiek, waarbij we het juridische kader, ethische vraagstukken en technische en statistische kenmerken van AI-systemen belichten.

Supervised learning toegepast in de zorg

Artificiële intelligentie functioneert op basis van twee fundamentele basiselementen: data en algoritmes. Data zijn de diverse informatiebronnen waarmee het systeem leert een output te genereren. Data kunnen bloedwaarden, pixels in een afbeelding of andere diagnostische parameters zijn. In de psychiatrie kunnen dit bijvoorbeeld resultaten van gestructureerde vragenlijsten zijn, aantekeningen uit patiëntdossiers, of gedragsobservaties. Algoritmes vormen het raamwerk van wiskundige regels en procedures die patronen in deze data identificeren, analyseren en gebruiken om voorspellingen mogelijk te maken.[2]

In het huidige medische landschap domineert *supervised learning* vanwege de controleerbare en reproduceerbare resultaten van deze techniek.[2,3] Deze voordelen zijn sterk afhankelijk van de kwaliteit en representativiteit van de trainingsdata.[4,5] Bij deze techniek krijgt een model meerdere gelabelde voorbeelden aangeboden – zowel de symptomen, metingen (*features*) als de uiteindelijke diagnose (*label*) die het moet voorspellen. Door het algoritme continu aan te passen en te optimaliseren verbetert het model zijn voorspellende waarde.

Een voorbeeld voor de psychiatrie zou kunnen zijn een AI-model dat getraind wordt om depressieve stoornissen te classificeren. Het model krijgt duizend geanonimiseerde patiëntdossiers aangeboden waarin specifieke symptomen zijn vastgelegd (zoals slapeloosheid, anhedonie, concentratieproblemen en stemming) samen met de uiteindelijke classificatie die door een regiebehandelaar is gesteld ('depressieve stoornis' of 'geen depressieve stoornis'). Het model leert patronen te herkennen: zo kan het

bijvoorbeeld detecteren dat een bepaalde combinatie van langdurige somberheid, slaapproblemen en gewichtsverlies een sterke indicator is voor de classificatie 'depressieve stoornis'. Bij een nieuwe patiënt met vergelijkbare symptomen kan het model dan de waarschijnlijkheid van een depressieve stoornis voorspellen. Naarmate het model meer voorbeelden verwerkt en feedback krijgt over de juistheid van zijn voorspellingen, worden de interne parameters in principe steeds verder verfijnd, wat leidt tot een nauwkeurigere classificatie. Echter, dit voorbeeld over een *supervised model* dat classificaties voorspelt, verdient een kritische en statistische nuancering.

Interbeoordelaarsbetrouwbaarheid van de DSM-5: relevantie voor trainen van AI-modellen

Interbeoordelaarsbetrouwbaarheid is een interessante maat in de geestelijke gezondheidszorg: ze laat bijvoorbeeld zien in hoeverre verschillende professionals tot dezelfde classificatie of diagnose komen. Dit wordt doorgaans gemeten met Cohens of Fleiss' κ, statistische methoden die geschikt zijn voor categorische beoordelingen en corrigeren voor toevallige overeenkomsten.[6] Met precies deze kappa-maten evalueerde men in de DSM-5-veldonderzoeken in 2012 de interbeoordelaarsbetrouwbaarheid van 23 diagnoses.[7] Van de uiteindelijk 15 geïncludeerde diagnoses bij volwassenen en 8 bij kinderen en adolescenten kregen 5 diagnoses een 'zeer goede betrouwbaarheid' toegekend (Cohens kappa = 0,60-0,79), 9 een 'goede betrouwbaarheid' (Cohens kappa = 0,40-0,59) en 6 een 'twijfelachtige betrouwbaarheid' (Cohens kappa = 0,20-0,39) en 3 een 'onaanvaardbaar bereik' (Cohens kappa = 0,20). Een kritische lezer kan stellen dat de auteurs van de veldonderzoeken hun terminologie zo afstemden dat de genoemde betrouwbaarheid beter lijkt dan de consistentie in werkelijkheid was.[7]

Aangezien we ons in dit artikel focussen op supervised-learning-modellen en LLM die behandelaars bij classificatie en diagnose ondersteunen of kunnen vervangen, kunnen we ons afvragen waarom deze uitkomsten relevant zijn. Hiervoor geldt: als Cohens kappa van de beschikbare data laag is, betekent dit dat de labels die gebruikt worden om een model te trainen, niet betrouwbaar genoeg zijn. De trainingsdata zijn statistisch van lage kwaliteit. Dit is een groot probleem voor het ontwikkelen van wiskundige regels, ofwel de AI-training. Het model moet zich baseren op gegevens die voor meerdere interpretaties vatbaar zijn en dus op zichzelf al onbetrouwbaar zijn. Uit de statistiek wordt Cohens kappa van > 0,6 en liever > 0,8 als noodzakelijk gezien om een betrouwbaar AI-model te trainen.[6,8] De nauwkeurigheid van onze classificatie en diagnoses voldoen daar grotendeels niet aan.

Er ontstaat dus een probleem als een AI-model wordt getraind op basis van het huidige psychiatrische classificatiemodel: het model vindt een patroon bij ongelijkwaardige input. Daarmee wordt het model dus beperkt door de inconsistentie van onze diagnostische labels. De lage Cohens kappa van het DSM-5-classificatiemodel weerspiegelt deze inconsistentie, en dat maakt supervised model learning in deze context technisch begrensd. Conclusie: supervised learning kan in dit domein nooit beter worden dan de betrouwbaarheid van de input waarop het is getraind – en die is laag. De betrouwbaarheid van het model is dus niet beter (maar ook niet slechter) dan die van de menselijke diagnostiek. En wanneer de betrouwbaarheid beperkt is, is de validiteit dat per definitie ook – zowel in AI als bij de huidige klinische praktijk.

EEN ETHISCHE BESCHOUWING

De voordelen van LLM's bij classificatie en diagnostiek

LLM's zijn geavanceerde AI-modellen die getraind zijn op enorme hoeveelheden tekst om taal te kunnen verwerken, genereren en bewerken. Toepassing van LLM's binnen de classificatie en diagnostiek kan verschillende voordelen bieden, zoals het verbeteren van de toegankelijkheid, objectiviteit, bijdragen aan meer gepersonaliseerde diagnostiek door de ondersteuning in verschillende talen, en het realiseren van kosteneffectieve diagnostiek.[9]

Generatieve AI-modellen zoals LLM's zijn net zo afhankelijk van juiste labels als predictieve AI zoals supervised machine learning. LLM's zijn al toegepast om risico's op bijvoorbeeld suïcide of gewelddelicten te detecteren door het analyseren van berichten op sociale media, en patronen te identificeren die traditionele methoden mogelijk missen.[10] Waarbij het label 'delict' of 'suïcide', in tegenstelling tot psychiatrische diagnostische labels, wél betrouwbare observaties oplevert en daarmee statistisch geschikt is voor artificiële toepassingen. Hoewel LLM's veelbelovend lijken voor de psychiatrie en geneeskunde in het algemeen, kennen ze ook risico's en roepen ze ethische vragen op over de rollen en verantwoordelijkheden bij het toepassen van deze diagnostische systemen.[11,12]

Leidt de inzet van LLM's mogelijk tot schade?

Confabulaties zijn onlosmakelijk verweven met de werking van grote taalmodellen: gedreven door statistische waarschijnlijkheden genereert het model de meest waarschijnlijke taalpatronen. Deze taalpatronen zijn niet per definitie feitelijk. Confabulaties zijn onjuiste antwoorden die wel overtuigend klinken.[13] Dergelijke onjuistheden kunnen ernstige gevolgen hebben, zoals het verstrekken van onjuiste medische informatie en het versterken van schadelijke stereotypen.[14]

Naast deze principiële kwestie zijn er ook implementatieproblemen. Zo kan overmatig vertrouwen in LLM's leiden tot verkeerde diagnoses of het uitstellen van professionele zorg.

Een andere bron voor mogelijke schade is te vinden in verschillende vormen van 'bias'.[9] LLM's zijn getraind op

grote datasets waarin bestaande vooroordelen worden weerspiegeld of zelfs vergroot. Dit kan leiden tot discriminatie, stereotypebevestiging of foutieve interpretaties van mentale gezondheidsproblemen.[9,15] Zelfs methoden om *bias* te verminderen bieden geen garantie dat schadelijke inhoud in de trainingsdata volledig wordt geëlimineerd.[15]

Daarnaast zijn de meeste modellen vooral getraind en voorzien van menselijke feedback in de Engelse taal, wat betekent dat deze modellen minder goed kunnen werken in andere talen.[15] Dit kan ongelijkheid qua toegang en kwaliteit van zorg vergroten, vooral bij migranten of minderheidsgroepen.

Ook brengen deze modellen privacyrisico's met zich mee, omdat ze vaak externe servers gebruiken en gevoelig zijn voor datalekken via technieken zoals '*prompt injection*'.[9] Bij deze techniek maakt men gebruik van een kwetsbaarheid in het model, waarbij misleidende of strategische invoer het model tot ongewenst gedrag kan aanzetten.[9] In de context van diagnostiek en classificatie, waar gevoelige informatie moet worden verwerkt, is databescherming extra belangrijk om schade en datalekken te voorkomen.

Welke vaardigheden zijn nodig voor verantwoord gebruik?

Voor verantwoord gebruik van AI-systemen, zoals LLM's voor diagnostiek of classificatie, is het van belang dat deze technologieën op een verantwoorde manier zijn ontwikkeld én dat gebruikers weten hoe ze deze dienen te gebruiken. Hun kennis en competenties kunnen de acceptatie en aanvaardbaarheid van LLM's bevorderen of ondermijnen: zelfs de best ontworpen technologieën kunnen immers tot schade leiden in de handen van iemand die ongeschoold is of die ze niet goed gebruikt.[16,20] De exacte vaardigheden zijn uiteraard afhankelijk van het type AI-toepassing, de toebedeelde rol, het type AI-systeem en binnen welke (sub)discipline het wordt toegepast. In de academische literatuur is nog weinig geschreven over de vaardigheden van specifiek *psychiaters* bij het gebruik van LLM's. In de literatuur over vaardigheden van artsen voor het verantwoord gebruik van medische AI worden de volgende zaken genoemd:

Technische en digitale vaardigheden. Hiertoe rekent men basiskennis van data-analyse, statistiek en informatiebeheer. Daarnaast is inzicht in AI-technologie, het kritisch beoordelen van AI-output en het opsporen van fouten essentieel voor verantwoord gebruik in de zorg. Verder wordt verwacht dat artsen kunnen samenwerken met AI- en datawetenschappers.[20] Specifiek voor grote taalmodellen is het belangrijk dat behandelaren weten hoe goede 'prompts' (opdrachten of instructies om AI aan het werk te zetten) gemaakt worden en wat de invloed van prompts is op de uitkomsten.

Intermenselijke en communicatieve vaardigheden. Door het gebruik van AI verwachten onderzoekers dat er meer ruimte komt voor en nadruk komt te liggen op 'kritische menselijke vaardigheden' die essentieel zijn voor klinisch redeneren en die AI niet (volledig) kan vervangen. Denk hierbij bijvoorbeeld aan emotionele intelligentie, empathisch luisteren, situationele beoordeling, interpretatie van vage signalen en toon, oogcontact maken en het signaleren van ongemak. Verder wordt de communicatie met patiënten, ook over het AI-systeem, gezien als belangrijke vaardigheid. Artsen dienen uitleg te kunnen geven over AI-gegenereerde informatie om gezamenlijke besluitvorming mogelijk te maken. Ten slotte moeten behandelaren omgaan met de veranderende toegang tot gezondheidsinformatie en LLM's, waarmee patiënten bijvoorbeeld zelf diagnoses stellen of hun gezondheid monitoren.[17]

Normatieve vaardigheden en deugden. Deze zijn belangrijk bij het gebruik van medische AI. Van behandelaren wordt bijvoorbeeld verwacht dat ze in staat zijn om door AI veroorzaakte schade te voorkomen, alert zijn op bias, het gestelde vertrouwen van de patiënt waarborgen en op basis van een veelheid aan informatie tot een diagnose kunnen komen.[16,17] Het betekent dat behandelaren moeten worden opgeleid om de grenzen van hun eigen kennis ('*epistemic humility*') te kennen én de grenzen van een AI-systeem te kunnen beoordelen.

Gebruik van LLM's en supervised machine learning: juridisch verstandig?

Los van de statistische inzichten en ethische aspecten is de vraag of het gebruik van *supervised machine learning en* LLM's juridisch gezien al mogelijk is in de zorg. Juridisch gesproken zullen AI's die in de zorg worden ingezet al snel gekwalificeerd worden als 'medisch hulpmiddel'. Daarvan is sprake indien ze vallen onder de (brede) definitie van een medisch hulpmiddel uit de Verordening Medische Hulpmiddelen (VMH).[18] Het gaat dan onder meer om software die door de fabrikant is bestemd om alleen of in combinatie te worden gebruikt bij de mens voor diagnose, preventie, monitoring, voorspelling, prognose, behandeling of verlichting van ziekte (art. 2 lid 1 VMH). Wanneer software wordt ingezet bij medische handelingen die (zeer) risicovol zijn, gelden daarvoor veiligheidsvereisten. In veel gevallen is dan keuring door een 'aangemelde instantie' vereist. Deze keuring resulteert doorgaans in een CE-markering. Sinds augustus 2024 is naast de VMH ook de AI-Verordening van toepassing binnen Europa.[19] Dit omvangrijke stuk wetgeving wordt in de komende jaren stapsgewijs uitgerold, en stelt onder meer regels aan het gebruik van 'hoogrisico-AI-systemen'. Als een AI-systeem een hoog risico heeft, zal het niet alleen waarschijnlijk CE-gemarkeerd moeten worden onder de regels van de VMH, maar gaan er ook verzwaarde vereisten gelden voor bijv. '*AI literacy*' door gebruikers van die systemen (artikel 4). Kort samengevat: gebruikers van AI-systemen die gebruikt worden voor bijv. diagnosticering, zullen – zoals we al omschreven – kennis moeten hebben van (de werking van) de systemen die zij gebruiken.

Dat is een lastig punt als we kijken naar de huidige '*state of the art*' van de in dit artikel beschreven systemen. Er bestaat ook zonder AI-systemen al heel veel variabiliteit in diagnosestelling door ggz-professionals. Maar met AI-systemen wordt dit duidelijk niet beter, zoals we in het voorgaande al beschreven. Bijkomend probleem daarbij is dat het voor gebruikers vaak niet te begrijpen is hoe AI-systemen tot hun diagnoses komen. Dat volgt uit het '*black box*'-karakter van de AI's. Maar vanuit het perspectief van een (medisch verantwoorde) onderbouwing van een diagnose is dat natuurlijk een probleem. Het probleem van hoe een AI tot antwoorden komt, speelt in het bijzonder bij vrij verkrijgbare LLM's, zoals ChatGPT en Claude. Die zijn immers níet ontwikkeld voor medische doeleinden, maar worden door professionals mogelijk wel daarvoor ingezet. Los van privacyvraagstukken (wat gebeurt er met de patiëntgegevens die worden ingevoerd?) geeft een medisch professional daarmee eigenlijk de diagnosestelling uit handen aan een stuk software dat daar niet voor getraind is. Dat is niet alleen problematisch in het licht van de AI-Verordening, maar ook vanuit de professionele verantwoordelijkheden zoals die voortvloeien uit de Wet Beroepen in de Individuele Gezondheidszorg (Wet BIG) en de Wet op de Geneeskundige Behandelingsovereenkomst (WGBO).
Met het laten keuren van nieuwe (medische) AI's zal hierin verbetering komen, en zorgprofessionals doen er goed aan op termijn alleen AI's te gebruiken die gecertificeerd zijn. Tot die tijd is het verstandig om voorzichtig (eigenlijk alleen in een experimentele setting) gebruik te maken van AI-systemen bij classificering in de ggz en zeer goed bedacht te zijn op de gevaren voor privacy van cliënten. Tevens is het goed te bedenken dat de input van een publiekelijk toegankelijke AI gebruikt wordt om het algoritme verder te trainen, wat tot (verdere) biases in het algoritme kan leiden.

CONCLUSIE

De inzet van AI in psychiatrische classificatie en diagnostiek biedt verleidelijke beloftes, maar er zijn op dit moment beperkingen.
Supervised learning en LLM's zijn afhankelijk van consistente labels, terwijl de interbeoordelaarsbetrouwbaarheid van de DSM-5 daarvoor simpelweg te laag is. Verder zijn LLM's (nog) niet ontwikkeld voor medische toepassingen, kunnen ze confabuleren, bias versterken en gevoelige data onbedoeld blootgeven. Het ontbreekt professionals bovendien vaak aan de noodzakelijke vaardigheden voor verantwoord gebruik. Tot slot is de inzet van niet-gecertificeerde AI in de zorg op dit moment juridisch niet zonder risico's.
Toch laten AI-systemen in andere domeinen binnen de ggz wél potentie zien – met name bij voorspellende modellen gebaseerd op objectieve uitkomsten, zoals wel of geen suïcidepoging, terugval of behandeluitkomst. Waar behandelaren vaak beperkt zijn in voorspellend vermogen, objectiviteit en snelheid, kan AI juist daar grote meerwaarde bieden. Daarnaast opent AI de deur naar een herijking van diagnostiek op basis van objectieve patronen in plaats van onze consensus over subjectieve classificaties.
We concluderen dat de technologie vooruitgesneld is op de praktijk. Voor nu blijft behoedzaamheid op zijn plaats. Niet alles wat technisch kan, is ook klinisch, ethisch of juridisch verantwoord.

De literatuurverwijzingen zijn online te raadplegen.

AUTEURS

Floriane Jaspers, kinder- en jeugdpsychiater en 'chief medical information officer', Levvel; eigenaar Kennisplatform AI in de GGz; onderzoeker ethiek en conversational AI, Amsterdam UMC.
Guido van Wingen, hoogleraar Neuroimaging in de Psychiatrie, Amsterdam UMC.
Karin Jongsma, universitair hoofddocent Bio-ethiek, Julius Centrum, Universitair Medisch Centrum Utrecht.
Roland Bertens, advocaat Gezondheidszorg, Van Benthem Keulen; universitair docent Gezondheidsrecht, UMC Utrecht.
Jeroen Legerstee, klinisch psycholoog i.o. en bijzonder hoogleraar Innovatieve behandeling bij kinderen en jongeren met dwang, angst en tics, Universiteit van Amsterdam en Levvel.

Correspondentie
Floriane Jaspers (Floriane.jaspers@levvel.nl).

Geen strijdige belangen gemeld.

Het artikel werd voor publicatie geaccepteerd op 25-6-2025.

Citeren
Tijdschr Psychiatr. 2025;67(8):473-476