Convergerende Diagnostiek: Naar een Convergerend Model van Mens en Machine in de GGZ

Deze bijdrage richt zich uitsluitend op het conceptuele en methodologische kader van het Convergerend Model. Juridische en regelgevende implicaties van implementatie in de praktijk blijven buiten beschouwing.

De technologie is op dit moment verder ontwikkeld dan de klinische praktijk. Kunstmatige intelligentie biedt mogelijkheden voor snellere, objectievere en efficiëntere diagnostiek. Echter wordt de toepassing van *supervised machine learning* en *large language models* (LLM's) in de geestelijke gezondheidszorg beperkt.

De reden dat dit nog niet grootschalig wordt toegepast komt door een fundamenteel probleem dat zich voordoet in het leren van deze systemen. LLM's en *supervised machine learning* leren op basis van de data die het gevoed wordt. Als de interbeoordelaarsbetrouwbaarheid (Cohens Kappa) van psychiatrische classificatiesystemen zoals de DSM-V al structureel laag zijn, dan ontbreekt het AI aan stabiele, eenduidige labels om van te leren. Om een betrouwbaar AI-model te trainen wordt een Cohens Kappa van minimaal .60 en liever >.80 als noodzakelijk gezien. (Jaspers et al., 2025; Landis & Koch, 1977; McHugh, 2012).

Het probleem dat Jaspers et al. (2025) schetsen is dat AI-modellen in de GGZ nooit beter kunnen worden dan de betrouwbaarheid van de input die het krijgt: een klinisch oordeel en classificatie op basis van de DSM-V. De betrouwbaarheid van het model is dus noch beter noch slechter dan huidige menselijke diagnostiek (Jaspers et al., 2025). Het ontbreek AI dus aan stabiele, eenduidige labels om van te leren en het is niet geheel onbekend dat er heel wat psychiatrische stoornissen zijn die symptomatisch sterk overlappen, en daarom een lage Cohens Kappa hebben. Zo is het onderscheid tussen depressieve stoornis, dysthyme stoornis en aanpassingsstoornis met depressieve stemming vaak ambigu, omdat duur en ernst van symptomen subjectief worden beoordeeld (Bayes et al., 2019; Zimmerman et al., 2015). De overlap illustreert dat deze diagnostische categorieën binnen de DSM niet altijd discreet zijn. Ze zijn grotendeels gebaseerd op klinische interpretatie en context.

De lage Cohens Kappa binnen de DSM-V (vaak <.6) vormt een belemmering voor zowel klinische consistentie als de ontwikkeling van betrouwbare AI modellen. Als professionals onderling sterk variëren in diagnostische beslissingen, wordt elke dataset waarop een algoritme wordt getraind blootgesteld aan ruis en interne inconsistentie. Dit betekent dat deze modellen niet betrouwbaar en valide kunnen leren wat bijvoorbeeld een

borderline persoonlijkheidsstoornis is, omdat de labels subjectief zijn. En een algoritme heeft objectieve, kwantitatieve data nodig om beslissingen te kunnen maken.

Een logische stap vooruit is dus niet om algoritmische modellen te verfijnen, maar het diagnostisch proces te herontwerpen. Een **simultane, dubbelblinde diagnostische cyclus** met een clinicus en algoritme kan een manier zijn om meer betrouwbare data te genereren. Dit levert drie cruciale voordelen op. Allereerst zou theoretisch gezien verbetering kunnen ontstaan van interbeoordelaarsbetrouwbaarheid. Door diagnoses parallel en onafhankelijk te laten ontstaan, kunnen discrepanties tussen beoordelaars en tussen mens en model zichtbaar en kwantificeerbaar worden. Het creëert een objectieve maat voor consistentie, waarmee niet alleen het AI-model, maar ook de menselijke diagnostiek iteratief verfijnd kan worden. Daarnaast ontstaat betere trainingdata voor AI. In een gesloten dubbelblind systeem convergeren menselijke beoordelingen na verloop van tijd richting een stabieler diagnostisch profiel. Het model leert niet meer van individuele willekeur, maar van gemiddelde consensus na gecontroleerde blindering. Dit verhoogt kwaliteit van labels en maakt trainingsdata voor AI statistisch betrouwbaarder. Tot slot ontstaat een versterking van de diagnostische betrouwbaarheid op individueel niveau. Door blind te werken (een algoritme en een clinicus), worden meerdere biases geminimaliseerd. Diagnostiek wordt zuiverder, transparanter en toetsbaar. Als AI-output namelijk systematisch afwijkt, kan dat worden gebruikt als feedbacksignaal. Waar mens en machine structureel niet overeenkomen, ligt vermoedelijk een blinde vlek in het classificatiesysteem zelf. Een gesloten, convergerende diagnostische cyclus biedt een methode waarin klinische expertise en kunstmatige intelligentie elkaar wederzijds versterken.

Referenties

Bayes, A., Parker, G., & Fletcher, K. (2019). Clinical differentiation of bipolar II disorder from borderline personality disorder. *Current Opinion in Psychiatry, 32*(1), 22–29.

Jaspers, F.A.F., Van Wingen, G.A., Jongsma, K.R., Bertens, R.M. & Legerstee, J.S. (2025) De rol van AI in diagnostiek en classificatie binnen de ggz. *Tijdschr Psychiatr*ie. 67(8):473-476

Landis JR & Koch GG. (1977). The measurement of observer agreement for categorical data. *Biometrics* 1977; 33: 159-74.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 276–282. https://doi.org/10.11613/bm.2012.031

Zimmerman, M., Chelminski, I., & Young, D. (2015). The frequency of DSM-5 diagnostic criteria for major depressive disorder. *Journal of Clinical Psychiatry, 76*(9), 1246–1251.