Zorgvraagzwaartemodel 1.0: naar een model van random zorgtoewijzing?

J. VAN OS

ARTIKEL

De ggz ondergaat een drastische transitie en wordt over de volgende jaren heringedeeld in drie domeinen: de huisartsen, bijgestaan door praktijkondersteuners voor de ggz (POH-GGZ); de generalistische basis-ggz; de specialistische ggz. Een belangrijk knelpunt in deze ontwikkeling is de vraag wie onder welk zorgonderdeel recht op zorg krijgt. Triage is deel van de optimalisering van de zorg.

In deze stelselwijziging speelt de huisarts een cruciale rol. Deze moet in alle gevallen de ggz-zorg indiceren. Hij of zij doet dit vanuit de zorgvraag van de patiënt, de gewenste expertise en de gewenste organisatie van de zorg rondom de patiënt. Voor een verwijzing naar de basis-ggz zijn er bijvoorbeeld 4 'producten' omschreven: basis kort (BK), middel (BM), intensief (BI) en chronisch (BC). Selectie (indicatie) gebeurt op basis van vijf aan zorgzwaarte gelieerde criteria: DSM-stoornis, ernst van de problematiek, risico, complexiteit, en beloop van klachten. Op dit ogenblik wordt er hard gesleuteld aan hulpmiddelen ('tools' in het jargon) om het verwijsproces te vergemakkelijken (en de zorg functioneler te maken).

In dit kader zag recent het eindadvies van de werkgroep Zorgvraagzwaarte (zvz) het licht, waarin de geboorte van *zorgvraagzwaartemodel 1.0* staat beschreven. Dit model, met nog een beetje sleutelen, moet dadelijk leidend worden bij de indicatie voor de zorg bij individuele patiënten en bij de onderhandelingen tussen zorgverzekeraars (inkoop) en zorgaanbieders (verkoop) bij zowel de basis-ggz als de specialistische ggz. Onder de leden van de commissie zien we twee mensen van Achmea, alsmede mensen van Menzis, Zorgverzekeraars Nederland, Stichting Benchmark GGZ, het ministerie van Volksgezondheid, Welzijn en Sport (VWS), de Nederlandse Vereniging voor Psychiatrie (NVvP), GGZ Nederland (GGZN) en het Landelijk Platform GGZ. De voorzitter is lid van de raad van bestuur van een ggz-instelling.

Welk probleem moet zorgvraagzwaartemodel 1.0 gaan oplossen?

Waar gaat *zorgvraagzwaartemodel 1.0* over? Kort gezegd, gaat het over enthousiaste, maar foutieve regressieanalysen van data. Op basis daarvan trekt men de volgende optimistische conclusie: '*de Werkgroep zorgvraagzwaarte is van mening dat een indicator die is opgebouwd uit deze variabelen zinvolle, gepaste en noodzakelijke informatie voor zorgvraagzwaarte bevat*'. De boodschap is duidelijk: hier wordt iets moois gebouwd, namelijk een lijstje dat hulpverleners dadelijk gaan invullen en waaruit gaat blijken tot welke zorgvraagzwaartecategorie elke individuele patiënt behoort. Hulpverleners zoals psychiaters en klinisch psychologen hebben hier blijkbaar moeite mee; het lijstje van de werkgroep Zorgvraagzwaarte (zvz) gaat hier verandering in brengen.

De lezer lijkt dit misschien onwaarschijnlijk, echter niet de werkgroep zvz – de regressiemodellen hebben het immers aangetoond. Hoewel de werkgroep terecht nog vele tekortkomingen en uitdagingen ziet, is de toon van het stuk gedetermineerd optimistisch: deze methodiek gaat werken, *no doubt about it*. Na *zorgvraagzwaartemodel 1.0* immers komt *zorgvraagzwaartemodel 2.0* en dan is het zo ver. Nieuwe lijstjes om in te vullen bij elke patiënt, wellicht weer op straffe van nieuwe boetes, waarna een computerprogramma van de zorgverzekeraar bepaalt wat de zorgvraagzwaarte van de patiënt in kwestie is. Softwarematige selectie bij de voordeur, als het ware.

Laten we verder kijken naar de regressiemodellen waar gewag van wordt gemaakt. Hiertoe moeten we naar een ander rapport, geheten: '*Verantwoording data-analyses in het kader van Werkgroep zvz 2012/2013*'. De auteurs hebben het stuk niet persoonlijk ondertekend – er staat slechts dat het rapport is opgesteld door DBC-Onderhoud in opdracht van Zorgverzekeraars Nederland en GGZN. Wat wordt er in dit rapport beschreven? Uit een bestand van ongeveer 1,25 miljoen behandeltrajecten in het landelijke DBC-informatiesysteem DIS worden er via allerlei stappen uiteindelijk 0,45 miljoen geselecteerd voor analyse. Vervolgens pro-

beert men in deze behandeltrajecten te voorspellen, per diagnostische hoofdgroep, wat een DBC duur maakt. Dit wordt uitgedrukt als (1) het aantal tijdgeschreven DBC-minuten, (2) het aantal ligdagen en (3) de productiewaarde (een variabele samengesteld uit behandeling en verblijf). Vervolgens probeert men de kosten te voorspellen met wat men (toevallig) voorhanden heeft in DIS aan variabelen, zoals wel of geen nevendiagnosen, score op de *Global Assessment of Functioning* (GAF), aanwezigheid van somatische en psychosociale factoren.

De analyse onder zorgvraagzwaartemodel 1.0 geanalyseerd

Een nadere studie van de resultaten die in het rapport worden gepresenteerd voor de psychotische stoornissen en het regressiemodel dat het aantal behandelminuten bij dit type stoornis probeert te voorspellen, levert boeiend materiaal. Het is immers bekend dat de psychotische stoornissen tot de 'duurste' behoren en dat er tevens enorme variabiliteit bestaat in zorgbehoefte en daarmee in benodigde behandelintensiteit. Een prospectieve zorgvraagzwaarte-index zou dus van harte welkom zijn.

Het eerste wat opvalt, is dat het regressiemodel statistisch gezien massaal overpowerd is; er zijn zoveel behandeltrajecten (meer dan 8000) opgenomen dat zelfs de miniemste verschillen in het model significant gaan uitvallen. Het heeft dus geen zin om het model te interpreteren in termen van statistische significantie – interpretatie moet gebeuren met betekenisvolle effectmaten zoals de mate van variantie in de afhankelijke variabele die door het model wordt verklaard (de zgn. *R-squared*).

Op deze manier bekeken, is het model teleurstellend: ondanks de overpowerde vergelijking vertonen maar zeer weinig variabelen in het model een significante associatie met de afhankelijke variabele. De *R-squared* is dan ook verwaarloosbaar klein, namelijk 1,8%. Met andere woorden: het model verklaart eigenlijk niets. Dit is een opmerkelijke bevinding, gezien de optimistische uitspraken van de werkgroep zvz.

Wellicht zijn psychotische stoornissen de uitzondering, en gaat het bij de andere stoornissen veel beter? Helaas, hier is het nauwelijks beter – hoewel er regelmatig sprake is van een of meer statistisch significante associaties in de massaal overpowerde modellen, blijft de *R-squared* verwaarloosbaar klein (in de orde van 2-4%, met een enkele uitschieter tot 6%). De werkgroep zvz echter blijkt alle resultaten uitsluitend te hebben geïnterpreteerd op het niveau van simpele statistische significantie, die enkel het gevolg is van de massale statistische overpower van de gebruikte modellen.

Bovendien begint nu iets anders op te vallen. De data in het DIS kunnen allerminst beschouwd worden als een homogene dataset – de structuur van de data bevat vele lagen van hiërarchische clustering, bijvoorbeeld per behandelaar, per afdeling en per instelling. De data analyseren zonder rekening te houden met deze klassieke lagen van hiërarchische clustering geeft verkeerde resultaten en maakt alle bevindingen in het rapport in feite onbetrouwbaar.

Laat ons samenvatten: we zien hier een groep beleidsmensen aan het werk die op basis van statistische significantie in massaal overpowerde analysen waar geen rekening is gehouden met velerlei bronnen van hiërarchische clustering uitspraken doen in de orde van '*de Werkgroep zvz is van mening dat een indicator die is opgebouwd uit deze variabelen zinvolle, gepaste en noodzakelijke informatie voor zorgvraagzwaarte bevat*'. De juiste uitspraak echter had moeten zijn: *de modellen laten zien dat de gebruikte variabelen van generlei waarde zijn met betrekking tot de zorgvraagzwaarte, gemeten in DBC-minuten, ligdagen of productiewaarde.*

Professionele ethiek en wetenschap botsen met wettelijke plicht

Het argument achter zorgvraagzwaarte 1.0 is dat indien men niet in staat is te prioriteren tussen goede en slechte zorg, de prestatiefinanciering onder druk komt te staan. En de zorgverzekeraar met lege handen achterblijft, zonder een valide systeem om zorg te indiceren, of om de klanten de best mogelijke zorg te garanderen. Tenzij er toch vertrouwen gegeven kan worden aan het oordeel van professionals, die hiervoor – althans voor psychiaters en klinisch psychologen – meer dan 10 jaar gestudeerd hebben en vaak hun zorg aanbieden in een multidisciplinaire context met interprofessionele controle. Hun registratie vraagt conti-

AUTEUR

JIM VAN OS, hoogleraar Psychiatrische Epidemiologie, vakgroep Psychiatrie en Psychologie, South Limburg Mental Health Research and Teaching Network, EURON, Maastricht UMC, Maastricht; King's College London, King's Health Partners, Department of Psychosis Studies, Institute of Psychiatry, Londen.

CORRESPONDENTIEADRES

Prof. dr. Jim van Os, vakgroep Psychiatrie en Psychologie, Maastricht UMC, Postbus 616 (locatie DOT12), 6200 MD Maastricht.

E-mail: j.vanos@maastrichtuniversity.nl

Geen strijdige belangen meegedeeld.

Het artikel werd voor publicatie geaccepteerd op 15-1-2013.

nue bijscholing en vergt dat professionals zich onderwerpen aan controle door vakgenoten. Weliswaar is fraude mogelijk, maar heeft men aangetoond dat een statistisch algoritme zoals de zorgvraagzwaarte 1.0 efficiënter is? Gaat een algoritme met random uitkomsten de klanten van de zorgverzekeraars toegang geven tot de kwalitatieve zorg waar ze volgens hun verzekering recht op hebben?
De aanlevering van patiëntendata (DBC, SBG, dadelijk zorgvraagzwaarteparameters) is de voorwaarde om het algoritmesysteem operationeel te maken. Hiertegen bestonden bezwaren. De minister heeft echter besloten een wettelijk kader te creëren om de aanlevering van data verplicht te stellen. Hiermee wordt iets wat wetenschappelijk laakbaar en deontologisch aanvechtbaar is, verplicht. De beroepsgroepen moeten zich bezinnen op de vraag of een wettelijk kader voldoende is om normen te schrappen. Iets wordt niet goed omdat er een wet is die het verplicht. Iets wordt ook niet kwalitatieve zorg omdat als je het niet doet, je geen geld meer krijgt. In het eerste geval primeert de (professionele) ethiek, in het tweede de wetenschap.

Het falen van zorgvraagzwaartemodel 1.0 was voorspelbaar

Maar wat zijn eigenlijk de variabelen die wél verschillen in zorgvraagzwaarte bepalen? Wat weten we daarover uit de literatuur, en wat betekent dat *a priori* voor de exercitie van de werkgroep zvz? Vreemd genoeg wordt hierover niets gezegd in het rapport. Men begint met regressiemodellen alsof het de eerste keer in de geschiedenis van de wetenschap is dat iets dergelijks wordt geprobeerd. Toch kunnen we beredeneren dat verschillen in DBC-behandelminuten en ligdagen in belangrijke mate afhankelijk moeten zijn van, onder andere, individuele verschillen en groepsverschillen in: (1) selectieprocessen op de filters van het pad naar de ggz die maken dat verschillende subgroepen terechtkomen bij verschillende instellingen en behandelaren met een specifiek profiel; (2) mate van respons op behandeling binnen deze instellingen en bij deze behandelaren; (3) verschillen in modificeerbare en niet-modificeerbare factoren die impact hebben op het natuurlijk beloop van de klachten; en (4) een heel scala aan procesvariabelen voor, tijdens en na de behandeling.
Ook weten we uit de literatuur dat zelfs als we al deze factoren nauwkeurig in kaart proberen te brengen, en te modelleren in predictieve modellen, de voorspellende waarde van het model, binnen een bepaalde diagnostische hoofdcategorie, zelden hoger wordt dan 20%. Dus net zoals bij SBG hadden de zorgverzekeraars de dure exercitie rond het bepalen van de zorgvraagzwaarte op basis van *dirty data* achterwege kunnen laten.
De verwachting van de werkgroep dat toevoeging van nog meer *dirty data* (namelijk die van Stichting Benchmark ggz) tot 'substantiële toename' van de voorspellende waarde gaat leiden, is ongegrond en getuigt van gebrek aan kennis en kunde op het gebied van predictief modelleren in het algemeen, en specifiek de literatuur hierover in de wetenschappelijke psychiatrie en psychologie. De '*member validity*' – het aantal mensen dat zich effectief gedraagt volgens de regels voorspeld door het algoritme – vastgesteld door een regressieanalyse, is klein. Een regressieanalyse kan vaststellen dat vrouwen gemiddeld 13 cm korter zijn dan mannen (167 versus 180), maar er zijn niet veel vrouwen die exact 167 cm lang zijn.

Primum non nocere en number needed to deny

Het rapport meldt expliciet dat de opdracht voor het bepalen van een zorgvraagzwaartemodel is opgehangen aan de wens van de zorgverzekeraars om dit te gebruiken voor het in- en verkoopproces van zorg, alsmede de bekostigings- en declaratiesystematiek in de ggz. Ondertussen heeft de minister in haar brief van 31 oktober 2013 de systematiek van zorgvraagzwaartebepaling, die in feite niet veel beter is dan random toewijzing, verplicht gesteld voor alle patiënten die van verzekerde zorg gebruik willen maken. Vooralsnog wordt onderkend dat het systeem niet bruikbaar is voor individuele indicaties – maar dat het dat in de toekomst wel zal worden. Het optimisme blijft, spijts de wetenschappelijke argumentatie dat dit onmogelijk is. Alleen: mogen we daar wel aan meewerken? Is random toewijzing aan een zorgvraagzwaartecategorie bij een patiënt met psychische klachten niet hetzelfde als een patiënt met tbc random laten kiezen uit een rijtje van 10 pillen, waarvan er slechts 1 genezend is voor zijn aandoening? Het moge duidelijk zijn dat random toewijzing in dezen schadelijk zal zijn voor de patiënt en botst met het principe van *primum non nocere*, dat besloten ligt in de hippocratische eed. Een professional staat dus in zijn of haar recht om medewerking met zorgvraagzwaartemodel 1.0 te weigeren. Iets voor de NVvP?
Maar laten we constructief zijn: op welke voorwaarden is meewerken met het zorgvraagzwaartemodel wél mogelijk? Laat ons hiertoe een nieuwe statistiek introduceren: het *number needed to deny* (NND). Het NND kan worden omschreven als: het aantal mensen dat een verkeerde zorgvraagzwaartecategorie krijgt toegewezen om 1 patiënt in de juiste (goedkope?) zorgvraagzwaartecategorie te plaatsen.
Een voorbeeld. We weten uit de literatuur dat als we *alles* nauwkeurig in kaart brengen met uitgebreide metingen en interviews, we in uitzonderlijke gevallen tot 20% van het beloop (en daarmee *by proxy* de behandelminuten) van een psychische stoornis, binnen een bepaalde diagnostische hoofdcategorie, kunnen voorspellen. De voorspellende waarde, of *post-test probability*, is dus 20%.

Stel, we proberen op basis van een *post-test probability* van 20% iemand tot de juiste categorie van goedkope zorgvraagzwaarte, passend bij dat beloop, toe te wijzen. Het *number needed to deny* is dan 5 – 1 = 4. Met andere woorden: zelfs bij de theoretisch hoogst mogelijke voorspellende waarde moeten er nog altijd 4 mensen aan de verkeerde zorgvraagzwaartecategorie worden toegewezen alvorens er één in de juiste categorie is geplaatst.

Een NND van 4 is duidelijk te hoog voor onze hippocratische eed. Als de voorspellende waarde stijgt tot 80%, daalt het NND tot 1. Dat betekent in dit geval dat om 1 persoon in de juiste categorie van zorgvraagzwaarte te krijgen, er tegelijkertijd een ander moet worden opgeofferd aan een onjuiste zorgvraagzwaartecategorie. Is dat wel acceptabel? Ik weet het niet – het *primum non nocere* betekent volgens mij wat anders.

De minister heeft beargumenteerd dat het zorgvraagzwaartemodel niet zal worden gebruikt voor *individuele* patiënten, maar om voor instellingen te classificeren hoe 'zwaar' hun patiënten *gemiddeld* zijn, zodat bij de zorginkoop hiermee rekening kan worden gehouden. Dit lost echter niets op: als het zorgvraagzwaartemodel 1.0 niet beter is dan random informatie, is het net zo min geschikt voor groepen als voor personen.

Is er een alternatief?

Er bestaat in Nederland, gegeven het latente geloof in marktwerking, geen planning of analyse van de capaciteit die de regionale ggz zou moeten hebben, zodanig dat er gewerkt kan worden volgens een zogenaamd *public health model* van zo goed mogelijke zorg voor zoveel mogelijk mensen met psychische zorgbehoeften. Dit is het model dat door de WHO wordt gehanteerd bij de verdeling van schaarse middelen om de gezondheid van de populatie te bevorderen. Hoewel een planning op het niveau van de populatiezorgbehoefte goed mogelijk is en de overheid aanzienlijk heeft geïnvesteerd om deze zorgbehoeften in kaart te brengen (NEMESIS-1 en -2), worden aanbod en capaciteit van de ggz in Nederland gedreven door historische parameters, gekenmerkt door (onverklaarde) regionale variatie in zowel het reguliere als het specialistische aanbod. Het is onbekend hoeveel onder- en hoeveel overbehandeling er plaatsvindt in de ggz en wat de rol is van regionale variatie in aanbod. Dit staat bekend als het 'Bestuurlijk Akkoord'.

De kerngedachte over de nieuwe ggz is de volgende: (1) valide wetenschap kan *vooraf* worden gebruikt om dingen logisch en beheersbaar in te richten op basis van het *public health model* van zo goed mogelijke zorg voor zo veel mogelijk mensen, (2) we stoppen met het systeem dat is gebaseerd op almaar uitbreidende, en uiteindelijk machteloze, controle en random zorgtoewijzing *achteraf*, (3) we gebruiken epidemiologische regio-indicatoren om de kwaliteit van de zorg te toetsen.

De constructiefouten in de systematiek van volumeafspraken kunnen nooit worden opgelost in het almaar uitbreidende paranoia-controle-registratie-afrekenklimaat dat de ggz is opgelegd. Beter het hele systeem op de schop en valide wetenschap gebruiken om oplossingen te vinden, dan schijnwetenschap aanwenden om staatspsychiatrie te creëren. Maar hierover meer in mijn boek dat eind maart verschijnt: *De DSM-5 voorbij! Persóónlijke diagnostiek in een nieuwe ggz.*

☞ Beide in dit stuk genoemde rapporten zijn te raadplegen op internet.

Het stuk van de werkgroep Zorgvraagzwaarte, Zorgvraagzwaarte ggz. Eindadvies werkgroep zorgvraagzwaarte. Definitieve versie (1.0): http://bit.ly/1eWlbsI

Het rapport van dbc-Onderhoud, Verantwoording dataanalyses In het kader van werkgroep Zorgvraagzwaarte 2012/2013 Versie: definitief 1.0: http://bit.ly/1hxiKPx