Wat ieder die betrokken is bij ROM zich over de metingen moet realiseren

A. HAFKENSCHEID, J. VAN OS

ACHTERGROND De kwetsbare aspecten die inherent zijn aan ROM-metingen binnen de ggz krijgen in de discussie over ROM opvallend weinig aandacht.

DOEL Gebruikers van ROM bewust maken van de beperkingen die aan ROM-metingen verbonden zijn.

METHODE Theoretische bespreking van drie kwetsbare aspecten die inherent zijn aan elk ROM-systeem in de ggz: arbitraire rekenregels; arbitraire meeteenheden en arbitraire constructen; geobjectiveerde meting van subjectieve informatie.

RESULTATEN De basis onder de metingen ten behoeve van ROM blijkt behoorlijk wankel.

CONCLUSIE Iedereen die bij ROM betrokken is, dient zich bewust te zijn van de meetproblemen die aan ROM verbonden zijn.

TIJDSCHRIFT VOOR PSYCHIATRIE 58(2016)5, 388-396

TREFWOORDEN meetproblemen, ROM-metingen

ARTIKEL

Sinds het themanummer over routine outcome monitoring (ROM) van 2012 vervult het *Tijdschrift voor Psychiatrie* in de wetenschappelijke forumdiscussie over ROM een voortrekkersol. In dit tijdschrift, maar ook elders, hebben wij ons de afgelopen jaren kritisch uitgesproken over de opzet en inrichting van ROM in de huidige vorm, waaraan verstrekkers van geestelijke gezondheidszorg in Nederland sinds enkele jaren verplicht zijn hun patiënten bloot te stellen (Hafkenscheid 2010, 2012, 2013; Van Os e.a. 2012; Hafkenscheid & Van Os 2013). Wij hebben verschillende alternatieven ter verbetering aangedragen (Hafkenscheid & Van Os 2014, 2015).

Onze kritiek richtte zich op een aantal ernstige zwaktes en tekortkomingen van het vigerende ROM-systeem: kritiekpunten die door Barendregt (2015), senior onderzoeker bij Stichting Benchmark GGZ (SBG), recent in dit tijdschrift werden gerelativeerd. Hij toont zich pleitbezorger voor handhaving van het huidige ROM-systeem. Zijn argumentatie voor handhaving komt erop neer dat ROM-metingen niet hoeven te voldoen aan de strenge psychometrische eisen die de wetenschap stelt, omdat de vergelijkingsinformatie voor benchmarken geen waarheidsvinding beoogt. Psychometrische kwaliteiten van monitorinstrumenten hoeven volgens hem niet meer dan 'goed genoeg' te zijn om te kunnen benchmarken voor 'best practices en kwaliteitsverbetering' (idem, p. 522).

Zo eenvoudig ligt het wat ons betreft niet. Het is absoluut waar dat psychometrische tekortkomingen minder zwaar wegen naarmate de vrijblijvendheid en relativiteit waarmee met uitkomsten van monitorsystemen wordt omgegaan groter zijn. Dat veronderstelt echter dat gebruikers voldoende kennis hebben van die vrijblijvendheid en relativiteit. Van gemiddelde gebruikers (clinici, beleidsmakers) kan niet verwacht worden dat zij over zulke kennis beschikken, dus ook niet dat zij zich de meet- en interpretatieproblemen van ROM-metingen erg bewust zijn.

Wij constateren dan ook dat er een diepe kloof gaapt tussen de onderliggende psychometrische theorievorming enerzijds en het gebruik van ROM in de praktijk anderzijds. Met imponerende formules en grafieken worden gemiddelde gebruikers gemakkelijk op het verkeerde been gezet. Door de uiterlijke en oppervlakkige gelijkenissen tussen metingen van fysieke en mentale attributen is het erg verleidelijk om de huidige ROM-systematiek in de ggz gelijk

te stellen met het monitoren van fysieke attributen of verschijnselen, zoals in de somatische geneeskunde gebeurt.

De huidige ROM-praktijk in de ggz wordt nogal eens retorisch verdedigd als 'eigenlijk de gewoonste zaak van de wereld', met verwijzingen naar de somatische geneeskunde, bijvoorbeeld het consultatiebureau. Niemand twijfelt aan het nut en de noodzaak van genormeerde groeidiagrammen die het consultatiebureau gebruikt om de lengte- en gewichtstoename bij jonge kinderen te monitoren, dus zou het monitoren van klachten, symptomen en probleemgedragingen in de ggz volgens sommigen (bijvoorbeeld Janssen e.a. 2013) net zo vanzelfsprekend moeten worden als het monitoren van somatische kenmerken. Zowel meettechnisch als inhoudelijk bezien gaat die vergelijking echter niet op.

Met dit artikel beogen wij betrokkenen bij ROM te helpen om meer kritische distantie te ontwikkelen ten opzichte van zowel de ROM-metingen als de analyse en interpretatie van deze metingen, los van het specifieke meetinstrument dat wordt gebruikt. We gaan achtereenvolgens in op drie kwetsbare aspecten die inherent zijn aan elk ROM-systeem in de ggz:

1. arbitraire rekenregels;
2. arbitraire meeteenheden en constructen;
3. geobjectiveerde meting van subjectieve informatie.

We vrijwaren onze bespreking zo veel mogelijk van jargon, om de leesbaarheid en toegankelijkheid voor de gemiddelde gebruiker van ROM te maximaliseren.

ARBITRAIRE REKENREGELS

Binnen de ROM zijn inmiddels twee statistische beslissingscriteria gangbaar om vast te stellen of een patiënt al dan niet heeft geprofiteerd van ggz-behandeling:

- de index voor betrouwbare verandering (voorgesteld door Jacobson en collega's: Jacobson e.a. 1984; Jacobson & Truax 1991);
- de index voor klinische significantie.

De *index voor betrouwbare verandering* (*reliable change*; RC) toetst of de scoreverandering tussen de voor- en de nameting de onbetrouwbaarheidsmarges van het monitorinstrument overstijgt. Hoe sterker het monitorinstrument onderhevig is aan meetfouten (bijvoorbeeld een klinische beoordelingsschaal met een lage interbeoordelaarsbetrouwbaarheid), des te groter de onbetrouwbaarheidsmarges waarmee we rekening moeten houden bij de interpretatie van scoreveranderingen tussen voor- en nameting. Waar we hier van 'nameting' spreken, kunt u ook 'vervolgmetingen' of 'follow-upmeting' lezen.

De *index voor klinische significantie* (*clinical significance*; CS) definieert de afkapwaarde voor de nameting, waarmee wordt gekwalificeerd of een behandeleffect 'klinisch betekenisvol' is. Deze index is op drie verschillende manieren geoperationaliseerd:

- Criterium a: een nameting (x_2, y_2) moet *buiten* het bereik van de klinische normpopulatie vallen, waarbij het bereik van de klinische populatie wordt gedefinieerd als twee standaarddeviaties (SD_k) van het gemiddelde voor de klinische normpopulatie (M_k), in de richting van de 'gezonde' normpopulatie.
- Criterium b: een nameting (x_2, y_2) moet *binnen* het bereik van de 'gezonde' normpopulatie vallen, waarbij het bereik van de 'gezonde' normpopulatie wordt gedefinieerd als twee standaarddeviaties (SD_g) van het gemiddelde voor de 'gezonde' normpopulatie (M_g).
- Criterium c: de kans dat een nameting (x_2, y_2) past binnen de 'gezonde' normpopulatie moet groter zijn dan dat deze past binnen de klinische normpopulatie.

Dit laatste criterium is inmiddels uitgegroeid tot conventie.

Problemen met deze rekenregels

De aantrekkingskracht van beide indexen is hun conceptuele eenvoud. De indexen zijn de pijlers onder het invloedrijke onderzoeksprogramma van Lambert en collega's (zie voor een overzicht: Lambert & Shimokawa 2011) naar de toegevoegde effectiviteit van feedback aan de therapeut over het scoreverloop op de *Outcome Questionnaire* (OQ-45) van elke behandelde patiënt. Beide rekenregels zijn echter niet zonder problemen. Het gevaar van reïficatie ligt al snel op de loer, terwijl de rekenregels tot op zekere hoogte arbitrair en artificieel zijn. Gemakkelijk kan uit het oog worden verloren hoe ongetoetst (en deels ontoetsbaar) de aannames zijn waarop deze indexen zijn gebaseerd. Met **FIGUUR 1** kan de logica achter de indexen eenvoudig worden gedemonstreerd, alsmede de interpretatieproblemen die de indexen kunnen opleveren.

AUTEURS

ANTON HAFKENSCHEID, klinisch psycholoog-psychotherapeut, Arkin/Sinai Centrum.

JIM VAN OS, hoogleraar Psychiatrische Epidemiologie, Maastricht UMC.

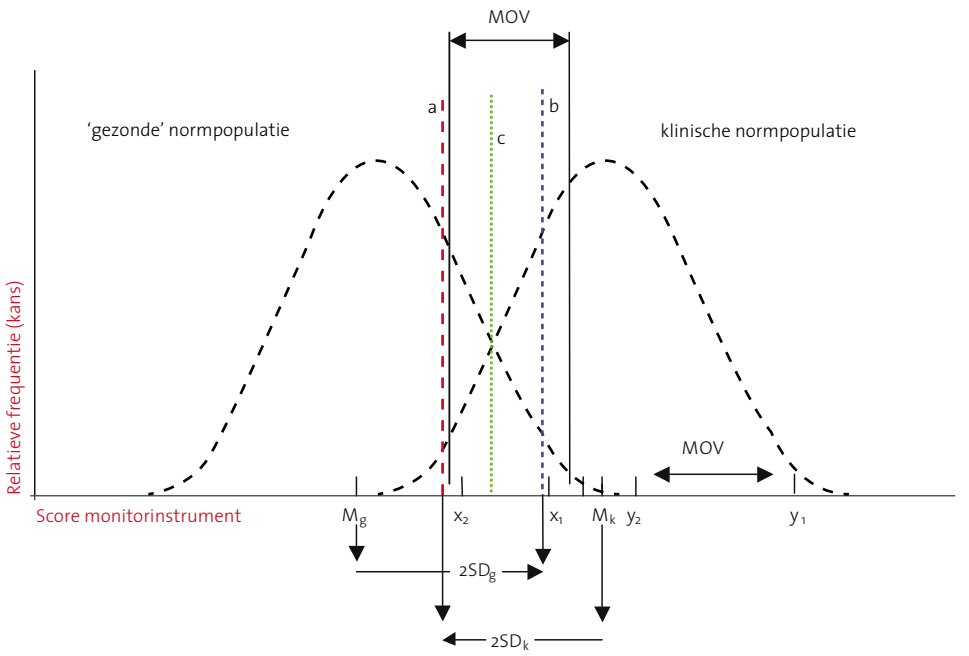
CORRESPONDENTIEADRES

A. Hafkenscheid, Arkin/Sinai Centrum, Joodse ggz, polikliniek Amersfoort, Berkenweg 7, 3818 LA Amersfoort.
E-mail: a.hafkenscheid@sinaicentrum.nl

Geen strijdige belangen meegedeeld.

Het artikel werd voor publicatie geaccepteerd op 11-11-2015.

FIGUUR 1 **Hypothetische scoreverdelingen monitorinstrument in de populatie**



M_g: gemiddelde 'gezonde' normpopulatie; M_k: gemiddelde klinische normpopulatie;

x_1: score voormeting patiënt x; x_2: score nameting patiënt x;

y_1: score voormeting patiënt y; y_2: score nameting patiënt y;

SD_g: standaarddeviatie 'gezonde' normpopulatie; SD_k: standaarddeviatie klinische normpopulatie;

MOV: marge onbetrouwbare verandering.

In **FIGUUR 1** staan de drie criteria voor klinische significantie weergegeven in de hypothetische populatieverdelingen voor patiënten (klinische normpopulatie) en 'normalen' ('gezonde' normpopulatie). Daarnaast is een hypothetische bandbreedte voor de onbetrouwbaarheid van het monitorinstrument in de figuur opgenomen. Tot slot zijn de scores voor (x_1, y_1) en na (x_2, y_2) behandeling voor twee hypothetische patiënten (x en y) in de figuur verwerkt. Beide patiënten scoren voor behandeling rechts van de afkapwaarden van alle drie de criteria voor klinische significantie, dus binnen de klinische normpopulatie.

Na behandeling scoort de patiënt x links van de afkapwaarden voor criteria b en c. Volgens deze criteria is deze patiënt dus 'klinisch betekenisvol' verbeterd. De scoreverandering van patiënt x valt echter binnen de bandbreedte voor onbetrouwbaarheid van het monitorinstrument. De scoreverandering is dus niet voldoende betrouwbaar en daarmee psychometrisch bezien zonder betekenis.

Patiënt y scoort na behandeling nog steeds rechts van alle drie de criteria voor klinische significantie, waarmee dus geen sprake zou zijn van klinisch betekenisvolle verandering. De scoredaling die deze patiënt laat zien, is echter groter dan die van patiënt x en deze scoredaling overstijgt wel ruimschoots de bandbreedte voor de onbetrouwbaarheid van het monitorinstrument. Een statistisch probleem is echter regressie naar het gemiddelde, waarbij extreem hoge scores (zoals in geval van patiënt y) bij de eerste meting op basis van toeval meer kans maken om substantieel lager uit te pakken bij een tweede meting.

In elk geval is het derde en gangbaarste criterium voor klinische significantie (criterium c) gebaseerd op de betwistbare en lastig toetsbare aanname dat patiënten en 'normalen' op populatieniveau weliswaar overlappende, maar duidelijk te onderscheiden scoreverdelingen vormen. Bij veel klinische klachten en problemen valt goed te verdedigen dat patiënten eerder een selecte (extremere) steekproef vormen uit de gehele bevolking.

Een andere betwistbare aanname is dat de scores op het monitorinstrument (voor patiënten en niet-patiënten) op zijn minst bij benadering normaal verdeeld zijn. Gemiddelde (M) en standaarddeviatie (SD) vormen uitsluitend geschikte parameters om de kansen te vergelijken dat een patiënt na behandeling meer in een klinische of 'gezonde' normpopulatie thuishoort als (bij benadering) aan de conditie voor een normale scoreverdeling in de populatie(s) is voldaan. Wanneer populatieverdelingen substantieel afwijken van de normaalverdeling zijn de mathematische eigenschappen van die scoreverdelin-

gen onzeker en is het domweg onmogelijk deze kansen te bepalen.

Een probleem van de index voor betrouwbare verandering is dat deze rekenregel weliswaar robuuste, maar erg conservatieve schattingen oplevert van de proportie betrouwbaar verbeterde of verslechterde patiënten binnen een cohort, zelfs bij monitorinstrumenten die zichzelf als behoorlijk betrouwbaar hebben bewezen (Wise 2004). Dat probleem geldt zeker wanneer een statistisch gangbare en kritische bandbreedte voor onbetrouwbaarheid (een symmetrisch, tweezijdig significantieniveau van 5%, gegeven de nulhypothese van onbetrouwbare verandering) wordt aangehouden. Uiteraard kan voor een minder kritische bandbreedte worden gekozen. Inherent aan die keuze is echter dat uitspraken over betrouwbare verandering dan noodzaken tot meer slagen om de arm (een grotere onzekerheidsmarge).

Een ander probleem van de index voor betrouwbare verandering is dat schattingen van de proportie betrouwbaar veranderde patiënten binnen een cohort sterk kunnen afhangen van de gebruikte schatter voor betrouwbaarheid (bijvoorbeeld interne consistentie of hertestbetrouwbaarheid), zelfs bij een monitorinstrument met over de gehele linie goede psychometrische eigenschappen (Hafkenscheid 1994).

ARBITRAIRE MEETEENHEDEN EN CONSTRUCTEN

Fysieke metingen: sterke kwantificeerbaarheid

Barrett (2003) geeft een brede definitie van wat meten eigenlijk inhoudt. Een representatieve theorie over meten stelt dat metingen moeten vastleggen hoe een empirisch relationeel systeem zodanig kan worden verenigd met een getallensysteem, dat iemand in staat is om hoeveelheden van een empirische entiteit te beschrijven aan de hand van die getallen. Het construct 'gewicht' is een klassiek voorbeeld van zo'n empirisch relationeel systeem. Als elke klasse objecten die gewicht als eigenschap heeft onderling kan worden vergeleken in relatie tot het kenmerk 'even zwaar als', dan kunnen de gewichten die zich in deze relatie tot elkaar verhouden aangemerkt worden als een relationeel systeem.

Theoretisch bezien is een vergelijkingsoperatie vereist tussen alle objecten in het systeem om vast te stellen: a. of de relatie stand houdt tussen willekeurig welke twee objecten, en b. of de relatie zoals die in abstracte zin is gedefinieerd in concreto kan worden waargenomen bij objecten die de eigenschap 'gewicht' krijgen toegekend. Voor de exacte wetenschappen is de representatieve theorie over meten buitengewoon vruchtbaar gebleken. Fysieke waarnemingen in de exacte wetenschappen kenmerken zich veelal door sterke kwantificeerbaarheid.

Wat betreft de vaststelling van behandeleffecten zijn er ook in de psychiatrie en de klinische psychologie wel voorbeelden van sterk kwantificeerbare metingen. Het aantal keren dat een patiënt met een obsessieve-compulsieve stoornis voor en na een gedragstherapeutische behandeling binnen een bepaalde tijdseenheid dwanghandelingen uitvoert, is zo'n voorbeeld. Andere sterk kwantificeerbare effectmetingen binnen de ggz zijn het aantal door een patiënt gepleegde delicten vijf jaar voor en vijf jaar na een forensisch psychiatrische behandeling of het aantal eetbuien per tijdseenheid voor en na behandeling van een eetstoornis. Zolang dergelijke metingen binnen het fysieke domein blijven, zijn ze niet-arbitrair.

Blanton en Jaccard (2006) waarschuwen ervoor dat sterke kwantificeerbare metingen binnen een niet-fysieke context onmiddellijk een arbitraire kwaliteit kunnen krijgen. Wanneer deze metingen worden gebruikt om psychologische constructen te operationaliseren, dan is er geen enkele garantie dat diezelfde metingen op psychologisch niveau nog steeds niet-arbitrair zijn. Deze auteurs noemen als voorbeelden de tellingen van nauwkeurig omschreven gedragingen of gebeurtenissen, zoals het aantal gerookte sigaretten per week of het aantal hoofdpijnaanvallen per maand. Als zodanig zijn deze tellingen niet-arbitrair, maar dat is wel het geval zodra ze worden opgevat als indexen voor psychologische attributen, zoals 'attitude tegenover roken' of 'stressgevoeligheid'.

Psychologische en psychiatrische constructen: zwakke kwantificeerbaarheid

Volgens Blanton en Jaccard (2006) is een fundamenteel probleem bij metingen van psychologische en psychiatrische constructen dat deze metingen veelal zijn gebaseerd op arbitraire meeteenheden. Zij noemen als voorbeeld een depressielijst die respondenten lokaliseert op de mate waarin zij zich depressief voelen op een meetschaal van 0 (de minimale score) tot 50 (de maximale score). De vooronderstelling bij de constructie van zo'n lijst is dat een respondent met een geobserveerde score van 35 depressiever is dan een respondent met een geobserveerde score van bijvoorbeeld 20.

De depressielijst geeft echter slechts een indirecte of afgeleide aanwijzing van de mate van iemands depressiviteit. De aanname is dat de geobserveerde scores op de depressielijst een eenduidige functionele relatie hebben met het onderliggende theoretische (klinische) construct 'depressie', dus met iemands 'ware' depressiviteit. De functionele relatie tussen geobserveerde score (de meting) en de onderliggende dimensie waarnaar deze score verwijst, is bij fysieke grootheden vaak evident, maar dat is bij psychologische dimensies zelden het geval. Uit een (geobserveerde) score op een depressielijst kan geenszins 'één-op-

één' de mate waarin iemand depressief is, worden afgeleid. Dat kan alleen wanneer de inhoudelijke (theoretische, praktische en klinische) implicaties van de geobserveerde depressiescores bekend zijn.

De in de psychologie en psychiatrie nog steeds dominante klassieke testtheorie is gebouwd op de axiomatische formule X = T + E. Daarin worden geobserveerde scores (X) gepostuleerd als samengesteld uit twee componenten: een (geschatte) 'ware score' (*true* score: T) en toevallige meetfout (*error*: E). Hoe groter T ten opzichte van E, hoe betrouwbaarder het meetinstrument. Geobserveerde scores X zijn perfect betrouwbaar in het theoretische geval dat E = 0. In dat geval is de geobserveerde score (X) identiek aan de 'ware score' (T).

Borsboom (2006) waarschuwt in dit verband voor een veelvoorkomende, maar ernstige misvatting: dat de 'ware score'-component ('waar' in de zin van 'vrij van meetfouten') van de geobserveerde score zou samenvallen met de 'werkelijke' score op het theoretische construct, aangenomen dat dit construct een lineaire kwantitatieve ordening kent. Veel onderzoekers kiezen uit pragmatisme, of op basis van conventie, voor de schijnoplossing van het 'blinde operationalisme': het hypothetische construct (de latente variabele) wordt domweg gelijkgesteld aan wat het meetinstrument meet.

Theoretische constructen, zoals depressie of angst, worden echter arbitrair als hun representatie wordt ingeperkt tot de somscore op een vragenlijst of beoordelingsschaal die geacht wordt het construct te meten, zonder goede causale theorie en zonder de 'surplusbetekenis' te vangen (Barrett 2003, 2005). Een voorbeeld is het gebruik van zelfrapportage om een emotie zoals angst te operationaliseren. Angst is een complex theoretisch construct, dat ten minste drie modaliteiten kent: een perceptie- of stimulusniveau, een semantisch of betekenisniveau en een niveau van responsprogramma's (Hermans e.a. 2007).

Barrett (2005) bespreekt de tekortschietende 'kruismodaliteit' (anders gezegd: de conditie dat elk van deze drie modaliteiten in de meting van het construct vertegenwoordigd is) in het angstonderzoek. De aanname dat de scores van een respondent op een zelfbeoordelingvragenlijst voor angst verwijzen naar een eendimensionale onderliggende variabele is empirisch onhoudbaar gebleken. Respondenten die op een monitorinstrument een laag angstniveau rapporteren in reactie op een stressinducerende empirische laboratoriumtaak kunnen fysiologische reacties tonen die juist indicatief zijn voor angst en omgekeerd.

'Blind operationalisme'

'Blind operationalisme' leidt tot cirkelredeneringen (Borsboom 2006): het te meten attribuut (theoretische construct) lijkt 'ontdekt' door de (gewogen of ongewogen) optelling van de items die voor de samenstelling van het meetinstrument zijn geselecteerd als de representatie van het attribuut op te vatten. De som van de items die geselecteerd zijn om het attribuut meetbaar te maken is echter niet meer dan een door de onderzoeker(s) geconstrueerde operationalisering van het attribuut. De 'wetenschappelijke steriliteit' die inherent is aan 'blind operationalisme' is principieel niet oplosbaar met het snel groeiende arsenaal aan geavanceerde en preciezere statistische en psychometrische analysetechnieken (Barrett 2005).

Michell (2000) levert wellicht de meest fundamentele kritiek op de meetcultuur binnen wetenschappen zoals de psychologie en de psychiatrie, die zich totaal niet (meer) lijken te bekommeren om de theoretische betekenis van de constructen die ze meetbaar proberen te maken. Volgens hem is die onverschilligheid in belangrijke mate het gevolg van een invloedrijk geworden, maar overmatig gesimplificeerde definitie van meten als 'het toekennen van cijfers aan objecten of gebeurtenissen volgens een bepaalde regel'. Elk construct is in principe meetbaar te maken, los van de vraag of en in hoeverre de geconstrueerde metingen zinvol of juist kunstmatig (of zelfs absurd) zijn. Dat leidt tot een vorm van zelfbedrog, die Blanton en Jaccard (2006) aanduiden als *meter reading*: onderzoekers en clinici kennen aan verkregen scores automatisch intrinsieke betekenis toe en projecteren de eigenschappen van sterk kwantificeerbare variabelen op de verkregen scores.

Terug naar Michell (2000): zodra er een instrument is ontworpen om het bedoelde, onderliggende construct te meten wordt simpelweg verondersteld dat het construct daarmee is opgehelderd of vastgelegd. De vraag in hoeverre onderliggende constructen zich voor kwantificering lenen, wordt overgeslagen en soms zelfs moedwillig genegeerd. Het is volgens hem op zijn minst aannemelijk dat veel constructen binnen de psychologie en psychiatrie hooguit op ordinaal niveau kunnen worden gemeten. Op ordinaal niveau is weliswaar een zekere rangordening tussen metingen mogelijk, maar de grootte van de verschillen tussen de proefpersonen of patiënten die op een bepaald attribuut zijn gemeten, kunnen niet zinvol worden bepaald en hebben derhalve geen betekenis. Sommige constructen zijn mogelijk zelfs geheel niet kwantificeerbaar.

GEFORCEERDE OBJECTIVERING VAN SUBJECTIEVE INFORMATIE

'Objectiviteit' van metingen in de somatische geneeskunde

Anders dan de ggz kent de moderne somatische genees-

kunde met name objectieve metingen, uitzonderingen daargelaten. 'Objectief' in de zin dat dataverzameling, dataverwerking en interpretatie van verkregen data gestandaardiseerd zijn, maar vooral ook 'objectief' in de zin dat het instrumentarium slechts beperkt afhankelijk is van zelfrapportage van de patiënt (of van idiosyncratische ideeën van de clinicus). Een voorbeeld van een uitzondering is de subjectieve pijnmeting: de arts die de patiënt vraagt om de ernst van zijn of haar pijn op een schaal van 0 tot 10 aan te geven.

Natuurlijk kennen de objectieve metingen in de somatische geneeskunde een zekere mate van onnauwkeurigheid (onbetrouwbaarheid) ten gevolge van imperfecties (subjectiviteit) van het expertoordeel (bijvoorbeeld: medisch specialisten of radiologen die een echo, röntgenfoto of MRI-scan verschillend interpreteren). Ook de grenswaarden (afkappunten) voor de scheidslijn tussen normale variaties enerzijds en 'verdachte' waarden (of regelrechte afwijkingen) anderzijds kunnen lokaal verschillen of door de tijd heen veranderen (bijvoorbeeld definities van overgewicht of hoge bloeddruk).

Om de gezondheidstoestand van de patiënt te monitoren beschikt de somatische geneeskunde niettemin over een heel arsenaal aan diagnostisch instrumentarium dat in meerdere betekenissen objectief is. Voor zover subjectieve zelfrapportage wel aan de orde is (bijvoorbeeld bij gehoortests of metingen van het gezichtsvermogen), zijn de beoordelingsprocessen die van de patiënt worden gevraagd heel eenvoudig en betrekkelijk eenduidig ('Deze letter is wat kleiner: Kunt u die nog lezen?').

In de somatische geneeskunde wordt de standaard (voor 'normaal' of 'afwijkend') grotendeels buiten de patiënt om bepaald. De patiënt vervult een passieve rol: hij *ondergaat* de somatische tests, zoals de bloedafname voor klinisch chemisch onderzoek. Los van de uitzonderlijke patiënten die hun testuitslagen moedwillig proberen te manipuleren (bijvoorbeeld verslaafden die niet hun eigen portie urine proberen in te leveren, maar die van een niet-verslaafde) kunnen de uitslagen van somatische tests maar zeer ten dele door de patiënt zelf worden beïnvloed.

'Objectiviteit' van metingen in de ggz

In de ggz heeft de 'objectiviteit' van metingen op een cruciaal punt doorgaans een veel beperktere betekenis dan in de moderne somatische geneeskunde: diagnostici en behandelaars zijn aangewezen op hun eigen subjectiviteit (het interpreteren van verbale uitspraken of gedragingen van de patiënt tegen de achtergrond van het eigen subjectieve referentiekader van de clinicus) en op de subjectiviteit van de patiënt (diens zelfrapportage van cognities, van emoties, van problemen buiten de therapeutische setting en van de oorzaak-gevolgrelaties van deze problemen).

De 'objectiviteit' van de zelfbeoordelingvragenlijsten of de klinische beoordelingsinstrumenten die in de ggz worden gebruikt, overlapt met de objectiviteit van het meetinstrumentarium uit de somatische geneeskunde wat betreft de gestandaardiseerde dataverwerking (algoritmes voor de optelling van antwoorden op items, etc.) en data-interpretatie (vergelijking van de scores van de patiënt met een referentiegroep die het meetinstrument eveneens heeft beantwoord). De data*verzameling* wordt echter noodgedwongen kunstmatig gestandaardiseerd: middels voorgeprogrammeerde items en meerkeuzevragen met vaste antwoordalternatieven (eventueel voorzien van welomschreven ankerpunten) wordt de subjectieve data-input kunstmatig geobjectiveerd. Het meten van mentale toestanden met zelfbeoordelingvragenlijsten of klinische observatieschalen in de ggz is dus niet meer dan isomorf met het meten van fysieke toestanden in de somatische geneeskunde.

Zelfbeoordelingvragenlijsten zijn in de ggz populair, omdat zowel de lijsten zelf als de patiënt (die als informant functioneert) veruit de gemakkelijkste en de goedkoopste informatiebron zijn. Het gebruik van zelfbeoordelinglijsten is ook inhoudelijk en ethisch goed verdedigbaar, omdat het subjectieve lijden van de patiënt vaak de belangrijkste component is van de behandeling.

Responstendenties, antwoordstijlen, interne standaarden

Zelfbeoordelingvragenlijsten zijn echter niet goed bestand tegen responstendenties of antwoordstijlen (bijvoorbeeld: de neiging om op elk item hoog of juist laag te scoren). Zelfbeoordelingvragenlijsten kunnen tevens sterk onderhevig zijn aan verwachtingseffecten bij patiënten, ook als zij die lijsten volledig naar eer en geweten invullen (wat meestal het geval is). Antwoordstijlen en verwachtingseffecten kunnen de validiteit van monitorinformatie ernstig ondergraven. Het is een illusie dat dergelijke vertekeningen gemakkelijk gecontroleerd kunnen worden, als ze al te traceren en in omvang te schatten zouden zijn.

Bij het invullen van zelfbeoordelingvragenlijsten worden van de patiënt als informant helemaal niet zulke eenvoudige afwegingen en overwegingen gevraagd, zelfs wanneer items en antwoordalternatieven eenvoudig en eenduidig zijn geformuleerd. Bij elk item waarin de patiënt gevraagd wordt aan te geven in welke mate zij/hij last heeft van een symptoom of klacht moet zij/hij een door haar/zijn leerervaringen gesynthetiseerde interne standaard aanspreken, om uit de antwoordalternatieven bij de items ('geen', 'nogal', 'heel veel') de meest passende te kunnen kiezen (Westen & Weinberger 2004). Deze interne standaard is het complexe (en in principe dynamische) product van zowel sociale vergelijking ('hoe somber voel ik mij ten

opzichte van mijn vrouw?'), als interne vergelijking ('hoe somber voel ik mij nu ten opzichte van toen ik mij nog niet had ziek gemeld?'). Het is alleen al om deze reden geenszins gezegd dat twee patiënten met exact dezelfde itemscores op precies dezelfde vragenlijst 'even depressief' zijn.

Verandering monitoren: verschuivende interne standaarden

De problemen die inherent zijn aan het gebruik van zelfbeoordelinglijsten bij de inventarisatie van klachten en symptomen bij aanvang van de behandeling stapelen zich verder op wanneer deze lijsten gebruikt worden om het verloop van de behandeling, dus om verandering, te monitoren. Dat laat zich het best illustreren met enkele voorbeelden, waarin het behandelverloop wordt gevolgd met *objectieve* monitorinformatie.

Een eerste voorbeeld van objectieve monitorinformatie is de gelopen afstand buitenshuis als uitkomstmaat bij een exposurebehandeling van een patiënt met agorafobie. Een tweede voorbeeld is het monitoren van gewichtsafname bij patiënten die een psychologische behandeling voor obesitas volgen. Een derde voorbeeld is de behandeling van dwangproblematiek: hoe vaak de patiënt over het verloop van de behandeling bij elk vertrek van huis controleert of de deur wel echt op slot zit.

In al deze drie voorbeelden (respectievelijk: toegenomen gelopen meters buitenshuis, gewichtsafname en afgenomen keren dat de voordeur wordt gecheckt) geeft de monitorgrafiek direct en meettechnisch correct het succes van de behandeling weer, ongeacht de vraag of de indicator zelf klinisch relevant is. De directe vergelijking van monitorscores op verschillende meetmomenten in de behandeling is geoorloofd, omdat de achtereenvolgende metingen zowel kwantitatief als kwalitatief vergelijkbaar zijn. Routinemonitoren heeft alleen zin zolang redelijkerwijs mag worden aangenomen dat de verschillende metingen over de loop van de behandeling dezelfde onderliggende meetschaal behouden, en dus vanuit eenzelfde betekeniskader kunnen worden geïnterpreteerd.

Die aanname is allerminst gegarandeerd wanneer monitorinformatie gebaseerd is op subjectieve zelfbeoordelingen. Het is op zijn minst plausibel dat de interne standaard van de subjectieve metingen voor de patiënt gaandeweg verschuift, bijvoorbeeld eenvoudigweg door feedback te geven op de monitorscores. Een verandering van perspectief op de eigen klachten, symptomen en interpersoonlijke problemen is zelfs vaak een therapeutisch doel of het geslaagde effect van de behandeling (Hofstee 1986). De keerzijde van een succesvol verschuivend perspectief is dat veranderingsscores over meerder meetmomenten hun vergelijkbaarheid of interpreteerbaarheid verliezen.

CONCLUSIES

In dit artikel hebben wij de focus gericht op slechts één van de draagbalken waarop ROM rust: de intrinsieke eigenschappen van ROM-metingen als zodanig. Deze wellicht belangrijkste draagbalk blijkt tamelijk gammel. Dat monitoren binnen de ggz methodologisch problematisch blijkt, is voor ons geen reden om monitoren van psychiatrische of psychologische (psychotherapeutische) behandelingen als een zinloze exercitie af te serveren. Het is geenszins gezegd dat metingen in de ggz nutteloos zouden zijn of weinig betekenis zouden hebben. Het punt is juist dat metingen in de ggz doordrenkt zijn van betekenissen. Die betekenissen zijn vaak gelaagd, ambigu en meervoudig en daarmee moeilijk generaliseerbaar over verschillende personen of over de tijd. Dat is geen enkel probleem zolang monitoruitkomsten 'sturend' worden gebruikt, namelijk om het therapeutisch proces te versterken en te beïnvloeden (Hafkenscheid 2014).

Sceptici zouden onze bespreking van drie basale problemen met meten binnen de context van de ggz kunnen opvatten als pleidooi om kwantitatieve metingen met monitorinstrumenten maar helemaal vaarwel te zeggen en alleen nog gebruik te maken van kwalitatieve informatie. Zo ver willen wij echter beslist niet gaan. Cijfers zijn binnen onze cultuur een steeds vanzelfsprekender onderdeel geworden van onze communicatie en vervullen in die zin een nuttige functie. Wij sluiten ons dan ook aan bij Barendregt (2015), voor zover hij het standpunt huldigt dat cijfers op zijn minst de communicatie tussen betrokkenen (patiënt en therapeut, behandelaars onderling) bevorderen. Kwalitatieve evaluaties horen een plek te hebben naast kwantitatieve metingen van behandeleffecten, maar hebben weer hun eigen methodologische moeilijkheden en interpretatieproblemen.

In de afgelopen decennia hebben de psychometrische en statistische analysetechnieken in de klinische psychologie en psychiatrie een explosieve ontwikkeling doorgemaakt, maar dat is voor de meettechnieken en -procedures zelf helaas niet het geval. Dat kan de klinische psychologie en psychiatrie niet echt verweten worden: de 'aard van de materie' waar de ggz zich mee bezighoudt, is het subjectieve beleven en dat is nu eenmaal een uitermate complex en inherent moeilijk grijpbaar fenomeen. Iedereen die betrokken is bij ROM - ongeacht of het de patiënt zelf, de behandelaar, onderzoeker of beleidsmaker betreft - zal zich er hoe dan ook voortdurend rekenschap van moeten geven dat analyses en interpretaties van ROM-metingen nooit en te nimmer sterker of krachtiger kunnen zijn dan deze metingen zelf.

LITERATUUR

- Barendregt M. Benchmarken en andere functies van ROM: back to basics. Tijdschr Psychiatr 2015; 57: 517-25.
- Barrett P. Beyond psychometrics: measurement, non-quantitative structure and applied numerics. J Manage Psychol 2003; 18: 421-39.
- Barrett P. What if there were no psychometrics?: constructs, complexity, and measurement. J Pers Assess 2005; 85: 134-140.
- Blanton H, Jaccard J. Arbitrary metrics in psychology. Am. Psychol 2006; 61: 27-41.
- Borsboom, D. The attack of the psychometricians. Psychometrika 2006; 71: 425-40.
- Hafkenscheid A. Rating scales in treatment efficacy studies: individualized and normative use. [proefschrift]. Groningen: Rijksuniversiteit Groningen; 1994.
- Hafkenscheid A. Rammelende ROM in de ggz: geen ROM zonder Routine Process Monitoring. GZ-Psychologie 2010; 2: 12-7.
- Hafkenscheid A. Subjectiviteit bij de interpretatie van het grafisch scoreverloop op monitorinstrumenten. Tijdschr Psychiatr 2012; 54: 29-134.
- Hafkenscheid A. Geen rad voor de ogen: reactie op De Jong & Van 't Spijker. Tijdschrift voor Psychotherapie 2013; 39: 203-7.
- Hafkenscheid A. De therapeutische relatie. Utrecht: De Tijdstroom; 2014.
- Hafkenscheid A, van Os J. Huidige ROM doet afbreuk aan valide kwaliteitsmeting. Tijdschr Psychiatr 2013; 55: 179-181.
- Hafkenscheid A, van Os J. Naar een deugdelijke ROM. MGv 2014a; 69: 20-28.
- Hafkenscheid A, van Os J. ROM van geïndividualiseerde behandeldoelen. PsychoPraktijk 2014b; 6: 29-32.
- Hermans D, Eelen P, Orlemans H. Inleiding tot de gedragstherapie. Houten: Bohn Stafleu van Loghum; 2007.
- Hofstee WKB. De begripsvaliditeit van retrospectieve voormetingen. Ned Tijdschr Psychol 1986; 41: 305-8.
- Janssen MMM, van Deurzen PAM, Klip H, Buitelaar JK. ROM in de kinder- en jeugdpsychiatrie: kansen en verplichtingen uitvoerbaar combineren. In: Buwalda VJA, Nugter MA, van Tilburg W, Beekman ATF, red. Praktijkboek ROM in de ggz II: implementatie en gebruik bij verschillende doelgroepen. Utrecht: De Tijdstroom. p. 41-7.
- Jacobson NS, Follette WC, Revenstorf D. Psychotherapy outcome research: methods for reporting variability and evaluating clinical significance. Behav Ther 1984; 15: 336-52.
- Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psych 1991; 59: 12-19.
- Lambert MJ, Shimokawa K. Collecting client feedback. In: Norcross JC, red. Psychotherapy relationships that work: evidence-based responsiveness. New York: Oxford University Press; 2011. p 203-23.
- Michell J. Normal science, pathological science and psychometrics. Theor Psychol 2000; 10: 639-67.
- Os J van, Kahn R, Denys D, Schoevers RA, Beekman ATF, Hoogendijk WJG, e.a. ROM: gedragsnorm of dwangmaatregel? Overwegingen bij het themanummer over routine outcome monitoring. Tijdschr Psychiatr 2012; 54: 245-53.
- Westen D, Greenberger J. When clinical description becomes statistical prediction. Am. Psychol 2004; 59: 595-613.
- Wise EA. Methods for analyzing psychotherapy outcomes: a review of Clinical Significance, Reliable Change and recommendations for future directions. J Pers Assess 2004; 82: 50-9.

SUMMARY

ROM measurements in mental health care: users need to be aware of the problems and pitfalls

A. HAFKENSCHEID, J. VAN OS

BACKGROUND The weaknesses inherent in ROM-data in mental health care are largely ignored in Dutch discussions about the pros and cons of ROM.

AIM To promote awareness among users and potential users of ROM with regard to the limitations of ROM data in mental health care.

METHOD We present a discussion of three types of measurement problems connected with the use of ROM data in mental health care: (a) arbitrary calculation rules for identifying changes that are reliable and of clinical significance, (b) arbitrary metrics and constructs and (c) forced objectivation of subjective information.

RESULTS ROM measurements are unreliable for use in mental health care because they lack a stable basis. The problems with these measurements are both psychometric and substantive.

CONCLUSION Anyone using or planning to use ROM measurements in mental health care should be aware of fundamental measurement problems associated with ROM.

TIJDSCHRIFT VOOR PSYCHIATRIE 58(2016)5, 388-396

KEY WORDS measurement problems, ROM data