De betrouwbaarheid van Delta-T

E. DE BEURS, L. WARMERDAM, J.W.R. TWISK

ACHTERGROND In de ggz wordt een felle discussie gevoerd over het nut van ROM en zinvolheid van benchmarken. Diverse opinies over de betrouwbaarheid en validiteit van prestatie-indicatoren passeren hierbij de revue.

DOEL Onderzoeken van de betrouwbaarheid van de voornaamste indicator van Stichting Benchmark GGZ (SBG), Delta-T, de indicator voor de behandeluitkomst.

METHODE De betrouwbaarheid werd met twee indices vastgesteld: de intraclasscorrelatiecoëfficiënt (ICC) voor de overeenstemming van herhaalde metingen van de gemiddelde behandeluitkomst en de rangordecorrelatiecoëfficiënt voor de consistentie in rangordening van zorgaanbieders over de tijd.

RESULTATEN Over het algemeen bleek de betrouwbaarheid van Delta-T uitstekend.

CONCLUSIE Betrouwbaarheid is een basaal vereiste, maar tegelijk slechts de eerste stap in het onderzoek naar de bruikbaarheid en zeggingskracht van Delta-T. Onderzoek naar de validiteit is volop gaande, met name naar hoe robuust de behandeluitkomst is voor vertekening door instrumentatie, selectie en casemixcompositie. De bruikbaarheid van de behandeluitkomst als indicator van kwaliteit van zorg zal zich uiteindelijk vooral in de praktijk moeten bewijzen.

TIJDSCHRIFT VOOR PSYCHIATRIE 60(2018)9, 592-600

TREFWOORDEN benchmarken, Delta-T, prestatie-indicator, ROM, test-hertest-betrouwbaarheid, therapie-uitkomst

ARTIKEL

Er is veel te doen over ROM en benchmarken in de ggz. Een van de zaken waarover volop wordt gediscussieerd, is de zeggingskracht van de therapie-uitkomstindicator van SBG, het gemiddelde pre-posttestverschil in klachten of functioneren rondom de behandeling dat is behaald bij een groep patiënten of de Delta-T. Delta-T is de indicator van het geaggregeerde resultaat van behandeling en wordt als een van de indicatoren van de kwaliteit van de geleverde zorg gebruikt (De Beurs e.a. 2017).

De Delta-T van een zorgaanbieder is het instellingsgemiddelde resultaat dat is behaald in een bepaald zorgdomein, bijvoorbeeld in de ambulante zorg voor veelvoorkomende psychiatrische stoornissen, in de zorg bij patiënten met ernstige psychiatrische aandoeningen of in de ouderenpsychiatrie. Delta-T wordt eerst per patiënt/behandeling berekend als de verschilscore tussen een voor- en nameting met een meetinstrument voor de ernst van de psychiatrische symptomen of het functioneren. Vervolgens kunnen scores worden geaggregeerd over groepen patiënten tot een prestatie-indicator, de gemiddelde Delta-T. Zo kan de Delta-T bepaald worden over alle behandelingen die in een bepaalde periode zijn afgesloten of over alle behandelingen bij een groep patiënten met een bepaalde stoornis.

Als prestatie-indicator voor een zorgaanbieder heeft Delta-T een praktisch bereik van 0 tot 15. Een Delta-T van 0 betekent dat de patiënten bij een zorgaanbieder gemiddeld niet zijn veranderd tijdens de behandeling, een Delta-T van 10 betekent dat de patiënten gemiddeld één standaarddeviatie zijn verbeterd.

De zeggingskracht van de prestatie-indicator Delta-T staat ter discussie, maar empirische gegevens over betrouwbaarheid en validiteit van Delta-T ontbreken vooralsnog. De eerste stap om de bruikbaarheid en de zeggingskracht van Delta-T empirisch te onderbouwen is de betrouwbaarheid van deze uitkomstindicator te onderzoeken. Betrouwbaarheid is een logische eerste stap, want betrouwbaarheid is een voorwaarde voor validiteit: een niet-betrouwbare meting kan niet valide zijn. Daarbij moet aangetekend

worden dat, zelfs wanneer het met de betrouwbaarheid goed zit, de validiteit van Delta-T ('Is dit de juiste indicator van de kwaliteit van geboden zorg?') nog niet is aangetoond. In de discussie komen we op dit punt terug.

De belangrijkste vereiste van een uitkomstindicator - of meetresultaat in algemene zin - is reproduceerbaarheid van het meetresultaat. Hierbij gaan we ervan uit dat de te meten entiteit hetzelfde blijft. Een voorbeeld: stel dat we het lichaamsgewicht willen bepalen. Wanneer twee verschillende weegschalen hetzelfde gewicht aangeven of wanneer dezelfde weegschaal bij herhaaldelijk meten telkens hetzelfde resultaat laat zien, dan is het aangegeven gewicht reproduceerbaar (en dus betrouwbaar) vastgesteld.

In de testtheorie staat deze vorm van betrouwbaarheid te boek als de test-hertestbetrouwbaarheid en het is de meest basale indicator van de psychometrische kwaliteit van een meetinstrument (Drenth & Sijtsma 1996). De test-hertest-betrouwbaarheid geeft aan in hoeverre het meetresultaat reproduceerbaar is wanneer we herhaaldelijk een stabiel kenmerk meten en het helpt ons om in te schatten of we het meetresultaat serieus kunnen nemen.

We hebben deze vorm van betrouwbaarheid onderzocht voor de Delta-T van zorgaanbieders in de ggz door herhaald hun Delta-T te bepalen en de uitkomsten te vergelijken (*agreement* over de tijd). Anders gezegd: als we ervan uitgaan dat de gemiddelde resultaten van zorgaanbieders over bijvoorbeeld een half jaar stabiel zijn, stellen we met deze vorm van reproduceerbaarheid vast of Delta-T ook hetzelfde blijft en dus een betrouwbare weergave biedt dan de werkelijke gemiddelde behandeluitkomst van een zorgaanbieder.

Twee vormen van reproduceerbaarheid worden onderscheiden (de Vet e.a. 2006): de *overeenstemming* tussen (herhaalde) metingen en de *consistentie* in het maken van onderscheid tussen te meten entiteiten. In de Engelstalige literatuur staan deze vormen bekend onder de naam '*agreement*' en '*consistency*'. Voor de reproduceerbaarheid van een meting wordt de intraclasscorrelatiecoëfficiënt (ICC) als indicator genomen (de Vet e.a. 2006); een ICC-waarde van 0 geeft geen reproduceerbaarheid aan, een waarde van 1 volmaakte overeenstemming. De consistentie in rangordening van zorgaanbieders wordt met Spearmans rangordecoëfficiënt *Rho* bepaald. Rho is de non-parametrische pendant van Pearsons correlatiecoëfficiënt; een waarde van 0 geeft aan dat twee rangordeningen ongerelateerd zijn, een waarde van 1 geeft perfecte overeenstemming in rangordening aan.

Een grafische manier om naar overeenstemming tussen metingen te kijken, die in de biostatistiek veel wordt gebruikt, staat bekend als de *limits of agreement*methode. Deze wordt geïllustreerd met plots volgens Bland-Altman (Bland & Altman 1986). In zo'n plot wordt voor alle zorgaanbieders het verschil tussen de herhaalde Delta-T's op de Y-as afgezet tegen de gemiddelde waarde van beide Delta-T's. Zo komt de omvang van het verschil aan het licht (de range tussen de limits of agreements of horizontale stippellijnen in de plots) en dat geeft aan hoe groot de willekeurige fout is tussen metingen. Ook worden eventuele systematische patronen zichtbaar, bijvoorbeeld dat verschillen toenemen naarmate de gemiddelde Delta-T hoger of juist lager is.

Resumerend: de overeenstemming van herhaalde metingen van de gemiddelde Delta-T van zorgaanbieders wordt dus op twee manieren vastgesteld: absolute overeenstemming (ICC) en consistentie in rangordening (Spearmans Rho); de overeenstemming wordt tevens grafisch inzichtelijk gemaakt met bland-altman-plots.

AUTEURS

EDWIN DE BEURS, hoofd wetenschappelijk onderzoek, Stichting Benchmark GGZ, Bilthoven, en hoogleraar ROM en Benchmarken, sectie Klinische Psychologie, Universiteit Leiden, Leiden.

LISANNE WARMERDAM, senior onderzoeker, Stichting Benchmark GGZ, Bilthoven.

JOS TWISK, hoogleraar Toegepaste Biostatistiek, sectie Methodologie en Toegepaste Biostatistiek, Vrije Universiteit, Amsterdam.

CORRESPONDENTIEADRES

Prof.dr. E. de Beurs, Faculteit Sociale Wetenschappen, Universiteit Leiden, Wassenaarseweg 52, 2333 AK Leiden.
E-mail: e.de.beurs@fsw.leidenuniv.nl

Strijdige belangen: prof. dr. Twisk meldde lid te zijn van de wetenschappelijke raad van SBG.

Het artikel werd voor publicatie geaccepteerd op 12-2-2018.

METHODE

Zorgdomeinen en meetinstrumenten

In de ggz worden verscheidene zorgdomeinen onderscheiden naar leeftijdscategorie (kinderen en jeugdigen, volwassenen en ouderen), naar doelstelling (cure of care) en naar aard (verslavingszorg is een apart zorgdomein). Voor dit onderzoek hebben we ons beperkt tot de ICC's van Delta-T in het zorgdomein Volwassenen Cure. Het gaat hierbij om ambulante zorg bij vooral stemmings- en angststoornissen van matige tot gemiddelde ernst. In het zorgdomein

TABEL 1 **Berekening van de intraclasscorrelatiecoëfficiënt (ICC) en de standaardmeetfout (SEM)**

De intraclasscorrelatiecoëfficiënt (ICC) wordt gedefinieerd als de variabiliteit (variantie) tussen zorgaanbieders gedeeld door de variabiliteit tussen zorgaanbieders + variabiliteit ten gevolge van systematische verschillen + de meetfout (alles tezamen de totale variantie) (de Vet e.a. 2006).

De ICC wordt berekend met ANOVA voor herhaalde metingen volgens een *two-way random effect*model (McGraw & Wong 1996). Als de meetfout klein is, dan benadert de ICC de waarde 1. Een veel geciteerde richtlijn om ICC waarden betekenis te geven is die van Cicchetti (1994): ICC < 0,40 is weinig betrouwbaar, 0,40-0,59 is redelijk, 0,60-0,74 is goed, ≥ 0,75 is uitstekend.

Uit de ICC kan de standaardmeetfout (*standard error of measurement*; SEM) berekend worden met de volgende formule: SEM = SD * √(1-ICC). Dit is handig, want zo wordt de ICC vertaald naar betrouwbaarheidsintervallen voor Delta-T op de oorspronkelijke meetschaal. Bijvoorbeeld: bij een ICC van 0,75 en een SD van 2,50 is de SEM 1,25 Delta-T punt. Het 95%-betrouwbaarheidsinterval is dan (1,96 * 1,25) ± 2,45.

Volwassenen Cure worden drie meetinstrumenten gebruikt om de behandeluitkomst te meten, de OQ-45 (Lambert e.a. 2004), de BSI (Derogatis 1975) en de SQ-48 (Carlier e.a. 2012). Dit zijn instrumenten met goede psychometrische eigenschappen en met een vergelijkbare responsiviteit (gevoeligheid voor verandering over de tijd) (Carlier e.a. 2017a; de Beurs e.a. 2012). Scores worden teruggebracht tot een uniforme schaal met een normale verdeling, de T-score. Dit is een meetinstrument onafhankelijke score met een gemiddelde bij de voormeting van $T_{pretest}$ = 50 (SD: 10) (De Beurs 2010).

Zorgaanbieders en behandelingen

In 2016 leverden 167 zorgaanbieders in Nederland maandelijks gepseudonimiseerde uitkomstgegevens aan bij SBG. Het betrof alle grote geïntegreerde instellingen, veel 'nieuwe' zorgaanbieders en een ruim aantal kleinere instellingen, zoals psychiatrische universiteitsklinieken en psychiatrische afdelingen van algemene ziekenhuizen; van bijna 90% van alle behandelingen in Nederland kwamen in 2016 gegevens binnen bij SBG. De eenheid van analyse voor dit betrouwbaarheidsonderzoek is de gemiddelde Delta-T van een zorgaanbieder voor alle initiële DBC's die werden afgesloten in de 1e en in de 2e helft van 2016. Om de Delta-T van een instelling mee te nemen in de analyse werd een minimaal aantal observaties van n = 25 evalueerbare initiële DBC's in elke periode vereist; bij < 25 observaties beschouwt SBG de Delta-T op voorhand als niet betrouwbaar.

Hierdoor hadden de uitkomsten betrekking op de wat grotere instellingen (n = 97), met minstens 25 evalueerbare behandelingen per half jaar. In totaal betrof het n = 56 212 initiële DBC's. Gemiddeld droegen deze 97 zorgaanbieders 585,6 (SD: 1115,7) evalueerbare DBC's bij aan de totale dataset; de range in het aantal DBC's per provider liep behoorlijk uiteen: van 55 tot 10.123. Het initiële DBC-deel van een behandeling is maximaal één jaar behandeling. Indien de behandeling langer dan een jaar duurt (50% van de behandelingen in het zorgdomein Volwassenen Cure), dan wordt een vervolg-DBC geopend en na weer een jaar (10% van de behandelingen) opnieuw. Vervolg-DBC's lieten we bij deze betrouwbaarheidsanalyse buiten beschouwing.

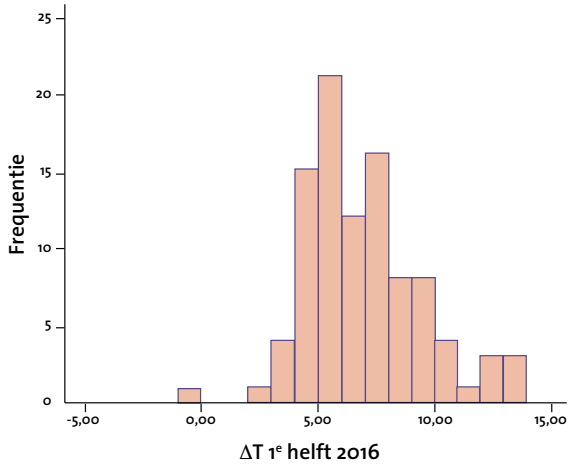
Statistische analyse

De intraclasscorrelatiecoëfficiënten (ICC's) van de Delta-T's van instellingen werden berekend, evenals de overeenstemming in rangordening van instellingen (Spearmans Rho) op basis van hun Delta-T. Dit werd gedaan voor alle behandelingen en voor behandelingen bij vijf subgroepen: patiënten met een stemmings-, angst-, somatoforme, persoonlijkheids- of ontwikkelingsstoornis (voornamelijk ADHD in de volwassenheid). Tevens bepaalden we nog eens apart de ICC en de Rho bij drie subgroepen patiënten die werden onderscheiden naar ernstniveau bij de voormeting: laag ($T_{pretest}$ < 45), gemiddeld (45 ≤ $T_{pretest}$ ≤ 55) en hoog ($T_{pretest}$ > 55). De berekening van de ICC en de standaardmeetfout (SEM) wordt beschreven in **TABEL 1**.

RESULTATEN

In **FIGUUR 1** is de frequentieverdeling te zien van Delta-T voor het initiële DBC van 97 instellingen in de 1e en 2e helft van 2016. De frequentieverdelingen waren normaal verdeeld met een gemiddelde van M = 6,96 (SD = 2,57) voor de eerste helft van het jaar en M = 7,02 (SD = 2,61) voor de tweede helft. De meeste instellingen hadden een Delta-T rond het gemiddelde; 21% van de instellingen had een Delta-T < 5,0, een aanzienlijk lagere waarde dan het landelijk gemiddelde; nog eens 16% van de instellingen had een

FIGUUR 1 **Frequentieverdeling van Delta-T-scores van 97 instellingen in de 1e en 2e helft van 2016**



Delta-T > 10,0, een aanzienlijk hogere waarde; de resterende 63% had een Delta-T van 5,0 tot 10,0.

TABEL 2 geeft de gemiddelde en standaarddeviatie van Delta-T's voor de 1e en 2e helft van het jaar, de betrouwbaarheidscoëfficiënten ICC voor overeenstemming en de standaardmeetfout SEM. Ten slotte laat **TABEL 2** ook de stabiliteit zien van de rangordening van instellingen op basis van hun Delta-T (Rho). In de rijen van **TABEL 2** staan deze gegevens voor alle patiënten en voor diverse subgroepen (diagnostische groepen en subgroepen onderverdeeld naar ernst van de klachten bij de voormeting).

Vergelijking 1e en 2e meting

Over de gehele groep was de Delta-T stabiel, maar bij een aantal subgroepen patiënten nam Delta-T significant toe over de tijd: bij stemmingsstoornissen ($F(1, 59) = 11{,}13$; $p < 0{,}001$), bij patiënten met een lage voormeting ($F(1, 47) = 6{,}07$; $p = 0{,}017$) en bij patiënten met een hoge voormeting ($F(1, 52) = 4{,}55$; $p = 0{,}038$).

Overeenstemming

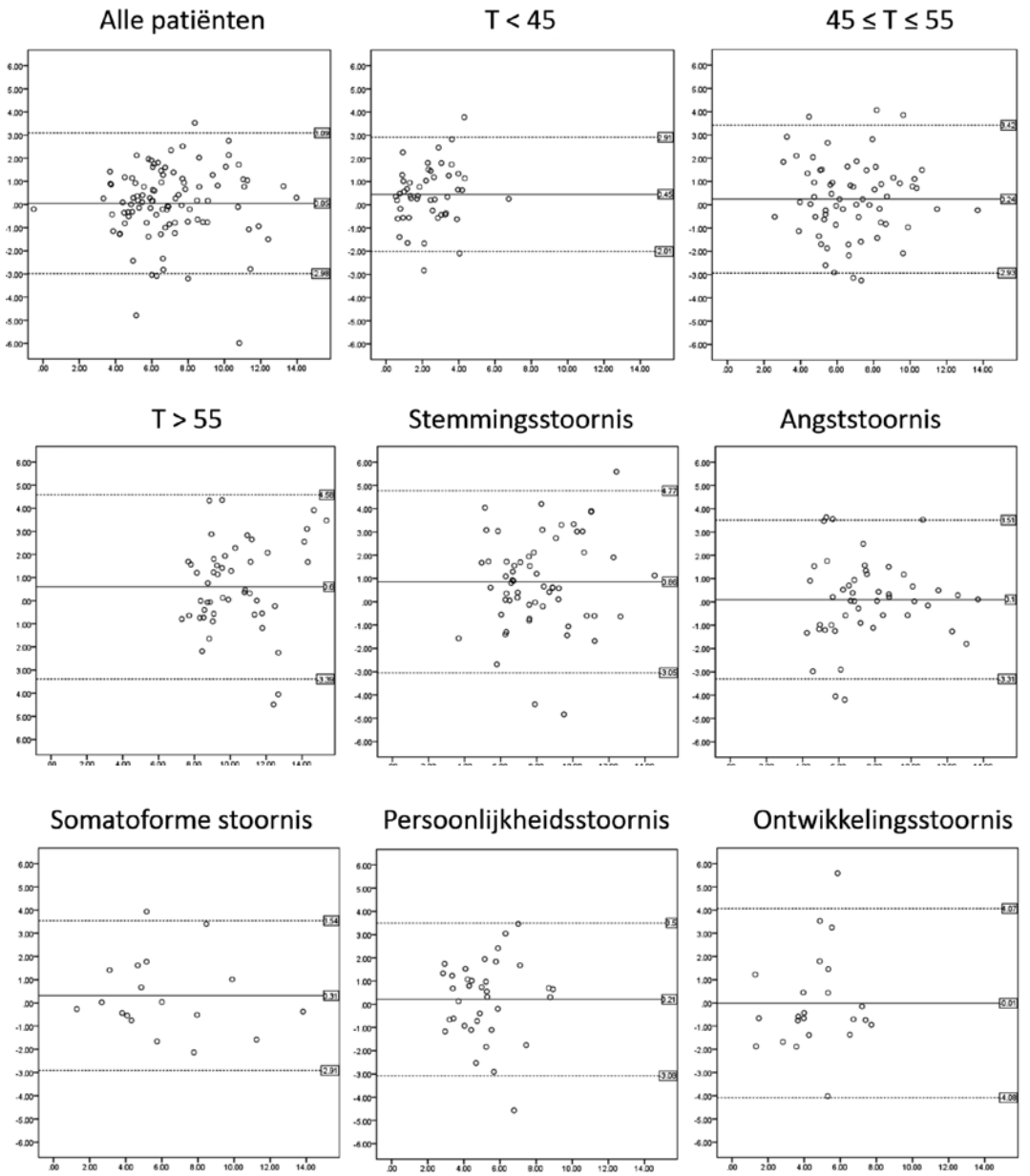
Voor de gehele dataset van beschikbare gegevens bleek Delta-T een zeer betrouwbare maat. Met een waarde van

TABEL 2 **Gemiddelde Delta-T in de 1e en 2e helft 2016 bij initiële DBC's van Volwassenen Cure, toets voor verschil, ICC's en Rhocorrelatie van rangordening**

		Delta-T Jan-juni		Juli-dec				
	N	M	SD	M	SD	ICC*	SEM	Rho**
Alle patiënten	97	6,96	2,57	7,02	2,61	0,82	1,10	0,78
Subgroepen:								
Lage voormeting (T < 45)	48	2,14	1,37	2,58	1,59	0,62	0,91	0,64
Gemiddelde (45 ≤ T ffi 55)	67	6,57	2,52	6,81	2,53	0,79	1,15	0,73
Hoge voormeting (T > 55)	53	10,57	3,01	11,16	2,99	0,76	1,47	0,70
Stemmingsstoornis	60	7,73	2,30	8,59	2,55	0,63	1,47	0,67
Angststoornis	48	7,52	2,56	7,62	2,63	0,78	1,22	0,73
Somatoforme stoornis	18	5,95	3,41	6,27	3,20	0,88	1,14	0,79
Persoonlijkheidsstoornis	36	5,07	1,83	5,29	1,86	0,59	1,18	0,57
Ontwikkelingsstoornis	23	4,64	1,99	4,63	2,25	0,53	1,45	0,43

N = aantal instellingen met minstens 25 observaties in elke periode; ICC = intraclasscorrelatiecoëfficiënt, ICC < 0,40 is weinig betrouwbaar, 0,40-0,59 is redelijk, 0,60-0,74 is goed, ≥ 0,75 is uitstekend (Cicchetti 1994); SEM = standaardmeetfout (standard error of measurement); Rho = Spearmans rangordecorrelatiecoëfficiënt, Rho: ≥ 0,70 is acceptabel (Streiner e.a. 2015).

FIGUUR 2 **Plots volgens Bland-Altman voor de relatie tussen de gemiddelde waarde van de Delta-T's voor de 1e en 2e helft van 2016 (op de x-as) en de verschilscore tussen deze Delta-T's (Delta-T_2 – Delta-T_1 op de y-as) voor alle patiënten, voor 3 ernstgroepen en 5 diagnostische subgroepen**



ICC = 0,82 werd de grens van 0,75 voor 'uitstekend' ruimschoots overschreden en kon Delta-T goed gebruikt worden om de uitkomsten van instellingen te vergelijken. De standaardmeetfout (SEM) bij deze betrouwbaarheid was 1,10 en het 95%-betrouwbaarheidsinterval lag ± 2,96 rondom de gevonden Delta-T-waarde.

De betrouwbaarheid nam af wanneer we naar subgroepen patiënten kijken. Met name bij stemmingsstoornissen leidde het verschil tussen de twee observaties van behandeluitkomst tot een lagere ICC en Rho. De betrouwbaarheid was echter nog steeds goed (7 van de 9 ICC > 0,60) en de toename van Delta-T doet zich bij alle instellingen in ongeveer gelijke mate voor. Bij persoonlijkheidsstoornissen en ontwikkelingsstoornissen (bij volwassenen, meestal ADHD) is de betrouwbaarheid redelijk.

De reproduceerbaarheid in rangordepositie van instellingen ten opzichte van elkaar (relatieve consistentie) is af te lezen uit de laatste kolom van **TABEL 2** (de Spearmans Rho-correlatiecoëfficiënten). Bij de Rho's zagen we hetzelfde beeld als bij de ICC's: over de gehele steekproef was de

betrouwbaarheid goed en bij subgroepen over het algemeen voldoende (bij van 5 van de 9 Rho ≥ 0,70). Ook bij stemmingsstoornissen kwam de rangordening van instellingen nog goed overeen. Bij twee subgroepen (persoonlijkheidsstoornissen en ontwikkelingsstoornissen) was de betrouwbaarheid lager.

Wanneer we naar overeenstemming bij de negen subgroepen keken met de plots volgens Bland-Altman (**FIGUUR 2**), dan bleek het verschil tussen de metingen voor de meeste instellingen klein, wat betekende dat de twee metingen van Delta-T een vergelijkbare uitkomst gaven. In de totale dataset van n = 97 lag het gemiddelde verschil dichtbij 0 en waren er slechts 6 instellingen waar het verschil tussen de metingen groter was dan ± 1,96 SD; ook in de andere dataset bleven de verschillen tussen herhaalde metingen overwegend binnen deze kritische grenzen. Voorts kwam uit de plots geen indicatie naar voren van een systematisch verband tussen het verschil (de positie op de y-as) en de gemiddelde waarde van de twee observaties (de positie op de x-as); met andere woorden: verschillen namen niet toe of af naarmate de Delta-T hoger of juist lager is.

DISCUSSIE

De resultaten wijzen uit dat Delta-T uitstekend betrouwbaar is en dat deze gebruikt kan worden om onderscheid te maken tussen instellingen of om veranderingen over de tijd te detecteren. De rangordening van zorgaanbieders is ook stabiel over de tijd. De bland-altmanplots laten voldoende overeenstemming zien tussen beide metingen en suggereren dat over het algemeen de verschillen binnen de grenzen van het betrouwbaarheidsinterval blijven en laten ook geen systematische vertekeningen zien bij zorgaanbieders met een hoge of juist een lage Delta-T.

Subgroepen

De betrouwbaarheid neemt enigszins af wanneer we naar subgroepen patiënten kijken. Bij 6 van de 8 onderzochte subgroepen is de betrouwbaarheid van Delta-T echter nog goed (ICC ≥ 0,60). Bij 2 subgroepen (persoonlijkheidsstoornissen en ontwikkelingsstoornissen, meestal ADHD bij volwassenen) blijft de betrouwbaarheid net onder de kritische grens van ICC = 0,60 en is dus hoogstens redelijk betrouwbaar. De reden voor het afnemen van de ICC is vooral dat bij deze twee groepen patiënten een lagere Delta-T wordt behaald en de spreiding in Delta-T tussen instellingen geringer is (verschillen tussen instellingen worden kleiner, zie de SD-waarden in **TABEL 1** en de spreiding op de x- y-as in **FIGUUR 2**). Bij deze stoornissen is de score een minder goede afspiegeling van de werkelijkheid en kan bijvoorbeeld een gevonden verschil tussen twee instellingen eerder een toevalsbevinding zijn. Ook is het bij deze stoornissen moeilijker om verschil te vinden in uitkomst tussen instellingen of over de tijd. Met een relatief onbetrouwbare indicator moeten er meer observaties verzameld worden om bijvoorbeeld een verschil tussen onbehandelde en behandelde patiënten aan te tonen.

De aanname dat Delta-T stabiel blijft over de tijd wordt door de bevindingen deels tegengesproken. Bij alle subgroepen neemt de Delta-T enigszins toe over de tijd. Bij patiënten met een stemmingsstoornis is de toename van Delta-T over de tijd het grootst. Bij initiële DBC's die in de tweede helft van het jaar worden afgesloten, wordt een gunstiger uitkomst behaald dan bij de DBC's die worden afgesloten in de eerste helft. Wellicht speelt hier een seizoensinvloed een rol (Winthorst e.a. 2014) en draagt een natuurlijk cyclisch patroon, met wat hogere symptoomscores in de winter en lagere in de zomer, bij aan dit verschil in Delta-T.

Vergelijking met andere meetinstrumenten

We weten nu dat we een betrouwbare indicator hebben. Een betrouwbaarheid van ICC = 0,82 is niet volmaakt, maar wel uitstekend en daarmee ruim voldoende betrouwbaar om onderscheid te kunnen maken, en ook betrouwbaar in vergelijking met andere indicatoren van kenmerken die we in de psychologie of psychiatrie meten.

Zo komt de betrouwbaarheid van psychiatrische diagnostiek in termen van overeenstemming tussen verschillende beoordelaars, zelfs als ze gebruikmaken van gestructureerde diagnostische interviews, niet boven kappa = 0,79 uit (Regier e.a. 2013): 9 van de 23 diagnoses die werden onderzocht in veldonderzoek ten behoeve van de DSM-5 hadden een betrouwbaarheid in de range van kappa = 0,40-0,59 (goed) en slechts 5 waren in de range van kappa = 0,60-0,79 (zeer goed).

Met de test-hertestbetrouwbaarheid van meetinstrumenten voor symptomen is het over het algemeen beter gesteld: Trajković e.a. (2011) rapporteren een ICC = 0,94 en ICC = 0,93 voor respectievelijk de test-hertest- en de interbeoordelaarsbetrouwbaarheid van de *Hamilton Depression Rating Scale*; Carlier e.a. (2017) vermelden voor de Nederlandse *Symptoms Questionnaire* (SQ-48) een test-hertestbetrouwbaarheid van ICC = 0,93 voor de totaalscore op dit zelfrapportage instrument.

Verder kijken dan Delta-T

Over het geheel genomen is het met de betrouwbaarheid van Delta-T dus goed gesteld. Betrouwbaarheid is echter slechts de eerste stap in het onderzoek naar de zeggingskracht van Delta-T als kwaliteitsindicator. We kunnen uw lichaamsgewicht betrouwbaar vaststellen met een weegschaal, maar of een reproduceerbaar vastgesteld gewicht ook iets zegt over uw algemene gezondheid is een andere kwestie. Daarvoor moeten ook andere factoren meegenomen worden, zoals uw

lengte. De *body mass index* (BMI) is al een betere indicator van een gezond gewicht dan alleen het lichaamsgewicht. Ook zegt het betrouwbaar vastgestelde aantal kilo's nog niets over de oorzaak van uw gewicht. (Onvoldoende of juist te veel calorieën tot u genomen? Een hoog metabolisme? Weinig beweging? Een ziekte onder de leden?)

Zo is het ook met de gemiddelde behandeluitkomst Delta-T van een zorgaanbieder in de ggz bij een bepaalde aandoening. Naar analogie met wat we hier stellen over lichaamsgewicht als indicator van algemene gezondheid is Delta-T alleen een te smalle basis voor een oordeel over de kwaliteit van een zorgaanbieder. Het wordt al beter als ook de lengte of de kosten van de behandeling meegenomen worden (De Beurs e.a. 2017b).

De betrouwbaarheid van Delta-T mag dan weliswaar uitstekend zijn, wat de gemiddelde Delta-T zegt over de effectiviteit, doelmatigheid of kwaliteit van de geboden zorg door een instelling is dus echter een andere kwestie. Een belangrijke factor die bijvoorbeeld nog niet is meegenomen, is de inspanning die een instelling levert om tot een bepaalde Delta-T te komen. Zo kunnen twee instellingen met exact dezelfde Delta-T sterk uiteenlopen in gemiddelde lengte of intensiteit van de behandeling.

Naar analogie met de bmi bij gewicht is het dus dienstig ook andere factoren mee te nemen bij de vergelijking. Zo zijn de behandelduur per Delta-T of de kosten per Delta-T wellicht een betere indicator voor de geleverde prestatie of de kwaliteit van een zorgaanbieder dan alleen Delta-T (De Beurs e.a. 2017b). Ook andere zaken dan symptoomreductie, zoals niveau van functioneren, de kwaliteit van leven of hoe de patiënt de behandeling ervaren heeft, zijn belangrijke indicatoren van kwaliteit van zorg, die nog buiten beschouwing zijn gebleven.

Verschillen tussen instellingen

Verder is nog niet opgehelderd waardoor de Delta-T wordt bepaald en wat mogelijke verklaringen zijn voor de aanzienlijke verschillen die we kunnen vaststellen tussen instellingen (range in Delta-T van 0 tot 15). Mogelijke oorzaken voor verschillen zijn:

– gebrek aan uniformiteit van meten;
– onvoldoende representativiteit van de data;
– confounding (Blijd-Hoogewys 2017).

Deze factoren verhogen de ruis in de data en beïnvloeden de validiteit van Delta-T in negatieve zin (Van Os e.a. 2012). Deze drie bedreigingen van de validiteit zijn de afgelopen jaren onderzocht. De vergelijkbaarheid van gegevens die afkomstig zijn van diverse meetinstrumenten blijkt beperkt (De Beurs e.a. 2012) en er kan 10-15% verschil in behandeluitkomst zijn vanwege gebruik van verschillende meetinstrumenten; na deze bevinding heeft SBG het aantal 'toegestane' meetinstrumenten teruggebracht.

Onderzoek naar de representativiteit heeft aangetoond dat bij voldoende ROM-respons (> 50% van de behandelingen evalueerbaar) de vertekening vanwege selectieve inclusie of selectieve uitval bij 95% van de instellingen niet groter is dan 0,5 Delta-T-punt; de kans op vertekening neemt toe naarmate de respons afneemt (De Beurs e.a. 2018). Onderzoek naar casemixvariabelen laat zien dat ernst van de klachten bij de voormeting de belangrijkste voorspeller is van de ernst bij nameting. Dit verklaart 25% van de variantie in nametingsscores; andere demografische en klinische casemixvariabelen verklaren samen nog zo'n 5% van de praktijkvariatie in behandeluitkomst (Warmerdam e.a. 2017). De SBG past statistische correctie toe voor confounding (casemixcorrectie).

Dit laat onverlet dat men goed moet blijven nadenken bij het vergelijken van behandeluitkomsten van verschillende instellingen. Andere bronnen van variatie in behandeluitkomst kunnen zijn de context waarbinnen een instelling werkt (verwijsbeleid of aantal zorgaanbieders in de regio) (Wennberg 2002). Verder kunnen er verschillen zijn in behandelwijze, in cultuur in de instelling, in hoeverre *evidence-based* gewerkt wordt volgens de nieuwste multidisciplinaire richtlijnen en zorgstandaarden, of men werkt volgens de principes van gedeelde besluitvorming (Metz e.a. 2015), in hoeverre men voortdurend ROM toepast bij de behandeling, enzovoorts (De Beurs e.a. 2017a).

Er is nog veel te onderzoeken in de ggz en gelukkig hebben we hiervoor met Delta-T een zeer betrouwbare indicator. Betrouwbaar vastgestelde verschillen tussen instellingen zouden nieuwsgierig moeten maken naar de achtergrond en oorzaak van deze verschillen. De gegevens als volstrekt nietszeggend terzijde schuiven is zonde en doet ze geen recht. Bovendien blijven mogelijkheden om de zorg in de ggz te verbeteren zo liggen en daarmee doen we onze patiënten te kort.

Ten slotte

Door regelmatig op de weegschaal te gaan staan val je niet af; daar zijn andere maatregelen voor nodig. Meten op zichzelf en delen van behandeluitkomsten (feedback) is dus niet voldoende; je moet ook met de meetresultaten aan de slag (De Beurs 2015). Organisatorische of behandelinhoudelijke maatregelen ter verbetering van de zorg in de ggz kunnen geïmplementeerd en geëvalueerd worden in een zogenaamde *plan-do-check-act*cyclus. Dit houdt kort gezegd in dat je een maatregel of interventie bedenkt (*plan*), implementeert (*do*), aan de hand van prestatie-indicatoren vaststelt of de maatregel effectief is ingevoerd (*check*) en vervolgens de maatregel evalueert en bijstelt (*act*). Zo'n maatregel kan een ingreep zijn in het proces ('de therapeutische relatie optimaliseren, behandelen volgens de richtlijn') of een organisatorische ingreep ('wachtlijst

bekorten door inzet van e-mental health' of 'regelmatig de behandeluitkomst bij de individuele patiënt evalueren en de behandeling op- of afschalen').
Voortdurend beschikbare feedback over eigen prestaties, het reflecteren op eigen resultaten en functioneren in alle geledingen van de instelling (individuele behandelaars, teammanagers, bestuurders van instellingen) en beproeven van diverse maatregelen ter verbetering van de zorg kunnen een betere ggz tot stand brengen.

LITERATUUR

- Beurs E de. De genormaliseerde T-score, een 'euro' voor testuitslagen. MGV 2010; 65: 684-95.
- Beurs E de, Barendregt M, Flens G, van Dijk E, Huijbrechts I, Meerding JW. Vooruitgang in de behandeling meten: Een vergelijking van vragenlijsten voor zelfrapportage. MGV 2012; 67: 259-70.
- Beurs E de. ROM en benchmarken, over meten, weten en wat dan? [oratie]. Leiden: Leiden University; 2015.
- Beurs E de, Barendregt M, Warmerdam L (red.). Behandeluitkomsten: bron voor kwaliteitsbeleid in de GGZ. Amsterdam: Boom; 2017a.
- Beurs E de, Warmerdam EH, Oudejans SCC, Spits M, Dingemanse P, de Graaf S, e.a. Treatment outcome, duration, and costs: A comparison of performance indicators using data from eight mental health care providers in the Netherlands. Admin Policy Ment Health 2017b; 44: 1-12.
- Beurs E de, Warmerdam EH, Twisk JW. Bias through selective inclusion and attrition: representativeness when comparing providers performance with routine outcome monitoring data. [submitted for publication 2018].
- Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 84767: 307-10.
- Blijd-Hoogewys E. Een introductie tot methodologische vraagstukken. In: de Beurs E, Warmerdam L, Barendregt M, red. Behandeluitkomsten: bron voor kwaliteitsbeleid in de GGZ. Amsterdam: Boom; 2017. p. 183-194.
- Carlier I, Schulte-Van Maaren Y, Wardenaar K, Giltay E, Van Noorden M, Vergeer P, e.a. Development and validation of the 48-item Symptom Questionnaire (SQ-48) in patients with depressive, anxiety and somatoform disorders. Psychiatry Res 2012; 200: 904-10.
- Carlier IVE, Kovács V, van Noorden MS, van der Feltz-Cornelis C, Mooij N, Schulte-van Maaren YWM, e.a. Evaluating the responsiveness to therapeutic change with Routine Outcome Monitoring: A comparison of the Symptom Questionnaire-48 (SQ-48) with the Brief Symptom Inventory (BSI) and the Outcome Questionnaire-45 (OQ-45). Clin Psychol Psychother 2017; 22: doi: 10.1002/cpp.1978.
- Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 1994; 6: 284-90.
- Derogatis LR. The Brief Symptom Inventory. Baltimore: Clinical Psychometric Research; 1975.
- Drenth P, Sijtsma K. Testtheorie: inleiding in de theorie van de psychologische test en zijn toepassingen. Houten: Bohn Stafleu Van Loghum; 1996.
- Lambert MJ, Gregersen AT, Burlingame GM. The Outcome Questionnaire-45. In: Maruish ME, red. The use of psychological testing for treatment planning and outcomes assessment. Volume 3: Instruments for adults (3rd ed). Mahwah: Lawrence Erlbaum; 2004. pp. 191-234.
- McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996; 1: 30-46.
- Metz MJ, Franx GC, Veerbeek MA, de Beurs E, van der Feltz-Cornelis CM, Beekman ATF. Shared Decision Making in mental health care using Routine Outcome Monitoring as a source of information: a cluster randomised controlled trial. BMC Psychiatry 2015; 15: 1-10.
- Os J van, Kahn R, Denys D, Schoevers RA, Beekman AT, Hoogendijk WJ, e.a. ROM: gedragsnorm of dwangmaatregel? Overwegingen bij het themanummer over routine outcome monitoring. Tijdschr Psychiatr 2012; 54: 245-53.
- Regier DA, Narrow WE, Clarke DE, Kraemer HC, Kuramoto SJ, Kuhl EA, e.a. DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses. Am J Psychiatry 2013; 170: 59-70.
- Streiner DL, Norman GR, Cairney J. Health measurement scales: a practical guide to their development and use. New York: Oxford University Press; 2015.
- Trajković G, Starčević V, Latas M, Leštarević M, Ille T, Bukumirić Z, e.a. Reliability of the Hamilton Rating Scale for Depression: A meta-analysis over a period of 49 years. Psychiatry Res 2011; 189: 1-9.
- Vet HC de, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol 2006; 59: 1033-9.
- Warmerdam L, Barendregt M, de Beurs E. Risk adjustment of self-reported clinical outcomes in Dutch mental health care. J Public Health 2017; 25: 311-9.
- Wennberg JE. Unwarranted variations in healthcare delivery: implications for academic medical centres. Br Med J 2002; 325: 961-4.
- Winthorst WH, Roest AM, Bos EH, Meesters Y, Penninx BW, Nolen WA, e.a. Self-attributed seasonality of mood and behavior: a report from the netherlands study of depression and anxiety. Depress Anxiety 2014; 31: 517-23.

SUMMARY

The reliability of Delta-T

E. DE BEURS, L. WARMERDAM, J.W.R. TWISK

BACKGROUND In Dutch mental health care there is an ongoing debate about the benefits of ROM and utility of benchmarking. Opinions vary regarding the reliability and validity of performance indicators.

AIM Investigation of the reliability of the main indicator of Foundation Benchmark Mental Health Care (SBG), Delta-T, the indicator of treatment outcome.

METHOD The reliability was established with two indices: the intraclass correlation coefficient (ICC) for the agreement between repeated assessments of average treatment outcome and the consistency in rank order of mental health care providers over time.

RESULTS The reliability of Delta-T proved to be excellent.

CONCLUSION Reliability is a basic requirement, but only the first step in establishing the utility of Delta-T. Further investigation of its validity is ongoing, especially on how robust treatment outcome is for bias due to instrumentation, selection, and confounding by casemix composition. Ultimately, the usefulness of treatment outcome as indicator of quality of care needs to be demonstrated in practice.

TIJDSCHRIFT VOOR PSYCHIATRIE 60(2018)9, 592-600

KEY WORDS benchmarking, Delta-T, performance indicator, ROM, test-retest reliability, treatment outcome