Tijdschrift voor
Gerontologie
en Geriatrie

DOI: 10.1007/s12439-013-0035-3
Gepubliceerd op: 07/08/2013

Wat is een reëel verschil bij herhaalde metingen met de Gedragsobservatieschaal voor de Intramurale Psychogeriatrie (GIP)?

Onderzoek bij deelnemers aan psychogeriatrische dagbehandeling

Auteurs: Han F. A. Diesfeldt

Samenvatting

De Gedragsobservatieschaal voor de Intramurale Psychogeriatrie (GIP) bestaat uit veertien subschalen die voldoen aan de voorwaarden van het Rasch-model. Zij meten verschillende aspecten van sociaal gedrag, cognitie en stemming. De GIP laat berekening van vier extra scores toe, voor hulpbehoevendheid, apathie, cognitie en affect. Doel van dit onderzoek was de reproduceerbaarheid van de GIP-scores vast te stellen. Op basis van beschikbaarheid werden 56 deelnemers aan psychogeriatrische dagbehandeling twee maal door dezelfde zorgverlener beoordeeld. Tussen beide beoordelingen verliepen gemiddeld (mediaan) 45 dagen (*interquartile range* 34–58 dagen). Reproduceerbaarheid werd bepaald door berekening van test-hertest *intraclass correlation* coëfficiënten ($ICC_{agreement}$). Kleinste betrouwbare (of minimale) verschilscores werden berekend op basis van de standaardmeetfout ($SEM_{agreement}$). De ICC's voor de achttien schaalscores varieerden van 0,57 (incoherent gedrag) tot 0,93 (angstig gedrag). Minimale verschilscores, gebaseerd op een 90%-waarschijnlijkheidsinterval, varieerden van 1 voor GIP-subschaal 14 (Angstig) tot 4 voor GIP-subschaal 12 (Somber), Hulpbehoevendheid en Apathie. Bij een 95%-waarschijnlijkheidsinterval varieerden de minimale verschilscores van 1 voor GIP-subschaal 14 (Angstig) tot 5 voor Hulpbehoevendheid. De resultaten ondersteunen de toepasbaarheid van de GIP voor de beoordeling van veranderingen in het gedrag van deelnemers aan psychogeriatrische dagbehandeling.

Interpreting change scores of the Behavioural Rating Scale for Geriatric Inpatients (GIP)

Abstract

The Behavioural Rating Scale for Geriatric Inpatients (GIP) consists of fourteen, Rasch modelled subscales, each measuring different aspects of behavioural, cognitive and affective disturbances in elderly patients. Four additional measures are derived from the GIP: care dependency, apathy, cognition and affect. The objective of the study was to determine the reproducibility of the 18 measures. A convenience sample of 56 patients in psychogeriatric day care was assessed twice by the same observer (a professional caregiver). The median time interval between rating occasions was 45 days (interquartile range 34–58 days). Reproducibility was determined by calculating intraclass correlation coefficients ($ICC_{agreement}$) for test-retest reliability. The minimal detectable difference (MDD) was calculated based on the standard error of measurement ($SEM_{agreement}$). Test-retest reliability expressed by the ICCs varied from 0.57 (incoherent behaviour) to 0.93 (anxious behaviour). Standard errors of measurement varied from 0.28 (anxious behaviour) to 1.63 (care dependency). The results show how the GIP can be applied when interpreting individual change in psychogeriatric day care participants.

Kernwoorden: affect, betrouwbaarheid, cognitie, dementie, gedragsobservatie, hulpbehoevendheid, sociaal gedrag, test-hertestdesign

Keywords: Affect, Care dependency, Cognition, Dementia, intrarater agreement, reliable change, Repeated measurements, social behaviour

Sinds 1987 kunnen psychologen in de ouderenzorg gebruik maken van de Gedragsobservatieschaal voor de Intramurale Psychogeriatrie (GIP).[1] De GIP kent 82 items, verdeeld over veertien subschalen, die op hun beurt ondergebracht kunnen worden in drie clusters: sociaal gedrag, cognitie en stemming.[2] De veertien subschalen beantwoorden aan de voorwaarden van het Rasch-model. Daarmee zijn de subschalen geschikt om reële individuele verschillen op de onderzochte gedragsdimensies aan het licht te brengen. De GIP wordt veel toegepast in verpleeghuizen bij diagnostische vragen rond cognitieve problemen en gedragsverandering.[3] GIP-observaties correleren met externe criteria zoals DSM-diagnose (dementie, amnestisch syndroom, psychose, of stemmingsstoornis) en met scores op andere gedragsobservatieschalen, zoals de Beoordelingsschaal voor Oudere Patiënten (BOP) en de Nurses Observation Scale for Inpatient Evaluation (NOSIE).[4,5] Het is vooralsnog onduidelijk in hoeverre de GIP geschikt is voor meting van veranderingen binnen personen over het verloop van de tijd.

Voor het vaststellen van een betrouwbare verandering is een psychometrische maatstaf nodig, ofwel een criterium waarmee beoordeeld kan worden of een scoreverschil tussen twee meetmomenten mag worden toegeschreven aan een reële verandering of aan meetfouten. De handleiding van de GIP geeft correlaties tussen gelijktijdige beoordelingen door telkens twee beoordelaars van 272 tot 274 patiënten.[1] De Pearsoncorrelaties varieerden van 0,53 (voor Angstig gedrag) tot 0,90 (voor Geheugenstoornissen). Betrouwbaarheidscoëfficiënten van 0,75 of hoger vermeerderen de nauwkeurigheid van een meting in verhouding tot de standaarddeviatie met 50% of meer.[6] Zeven subschalen van de GIP voldoen aan dit criterium.[1] De gegevens zijn echter verkregen bij bewoners van verpleeghuizen en daarmee niet zonder meer toepasbaar voor deelnemers aan een psychogeriatrische dagbehandeling. Bovendien is een Pearsoncorrelatiecoëfficiënt niet de meest geschikte maat voor de betrouwbaarheid van herhaalde metingen.[6,7]

Er is voor de GIP verder geen onderzoek bekend met herhaalde metingen. Voor enkele andere geriatrische gedragsobservatieschalen, zoals de BOP, de Neuropsychiatric Inventory (NPI) en de Cohen-Mansfield Agitation Inventory (CMAI) zijn dergelijke gegevens wel beschikbaar, zij het dat het onderzoek betrekking had op bewoners van verpleeghuizen, niet op deelnemers aan psychogeriatrische dagbehandeling.[8,9,10,11] Het onderzoek liet zien dat soms aanzienlijke scoreverschillen nodig zijn om te kunnen spreken van een daadwerkelijke gedragsverandering.

Voor de klinische praktijk is het van groot belang te weten welke betekenis kan worden toegekend aan verschilscores binnen een en dezelfde patiënt. In dit onderzoek bij deelnemers aan psychogeriatrische dagbehandeling zijn herhaalde metingen uitgevoerd, waarbij dezelfde zorgverlener patiënten met een tussenperiode van enkele weken twee keer beoordeelde.

Onderzoeksopzet

Het onderzoek werd uitgevoerd in twee centra voor psychogeriatrische dagbehandeling. Patiënten bezoeken de dagbehandeling twee keer per week. Anderhalve maand na aanvang van de dagbehandeling vult een vaste medewerker (activiteitenbegeleider en zorgplancoördinator) de GIP in. Vervolgens gebeurt dat periodiek, ter voorbereiding van de halfjaarlijkse zorgplanbesprekingen. De tamelijk complexe berekening van GIP-subschaalscores werd elektronisch uitgevoerd, met behulp van Excel. De software berekende ook de scores voor de drie clusters Apathie (9), Cognitie (9) en Affect (10). Tussen haakjes het aantal GIP-items waarop de clusterscores berusten.[12] Tevens werd op basis van tien items uit de GIP een score voor hulpbehoevendheid berekend.[5]

Van november 2009 tot december 2011 werd elke achtereenvolgende patiënt voor wie een GIP werd ingevuld opgenomen in een hertestbestand. Dat gold zowel voor deelnemers die zes weken eerder met dagbehandeling begonnen waren, als voor deelnemers die al langer in dagbehandeling waren. Drie tot vier weken later vroeg de onderzoeker aan de medewerker die de GIP had ingevuld om dit voor dezelfde patiënt nog eens te doen. Daarbij werd toegelicht dat het onderzoek niet was bedoeld om na te gaan hoe 'betrouwbaar' de medewerker als beoordelaar was, maar om te onderzoeken met welke spontane fluctuaties rekening gehouden moet worden om verschillen over het verloop van de tijd te interpreteren. De betrokken

medewerkers waren ieder verantwoordelijk voor een kleine groep patiënten en hadden minstens twee jaar ervaring met het invullen van de GIP. Het beoogde interval van drie tot vier weken tussen eerste en tweede invulling van de GIP werd kort genoeg geacht om de kans op werkelijke gedragsveranderingen klein te houden en lang genoeg om een gedragsregistratie te verkrijgen die op nieuwe, oorspronkelijke waarnemingen zou berusten.

Gevraagd werd de tweede GIP in te vullen naar de bevindingen van de afgelopen twee weken, zonder raadpleging van eerder ingevulde GIP-formulieren. Als de medewerker weet had van een serieuze verandering in de toestand van de patiënt (zoals ziekte in de periode na de eerdere invulling van de GIP, ernstige gebeurtenissen thuis of in de familiesfeer, langdurige afwezigheid, of opvallend veranderd gedrag), werd de herhalingslijst oningevuld aan de onderzoeker teruggegeven.

Variantiebronnen

Volgens deze opzet (herhaalde metingen door dezelfde beoordelaar) zijn er drie variantiebronnen om in de betrouwbaarheidscoëfficiënt te verdisconteren. Naast verschillen tussen de beoordeelde patiënten zijn dat verschillen op de twee registratiemomenten en ten slotte de variantie als gevolg van (niet-systematische) meetfouten. Gesteld dat patiënten tussen de gekozen meetmomenten niet veranderen, zijn er toch verschillen tussen de twee registratiemomenten te verwachten omdat dezelfde beoordelaar de eerste keer misschien meer of minder nauwkeurig registreert dan de tweede keer, of vanwege kleine dagdagelijkse veranderingen in het gedrag van de beoordeelde patiënten. Dit laatste is vooral te verwachten bij het registreren van stemmingsvariabelen, waarvoor de GIP diverse observatieschalen bevat.

Variantieanalyse

Voor het berekenen van het aandeel van elk van de drie variantiebronnen (patiënten, meetmomenten en meetfouten) is voor de GIP-scores een variantieanalyse voor herhaalde metingen gebruikt (via de SPSS-routine Reliability). Op basis van de drie variantiecomponenten is een *intraclass correlation coefficient* (ICC) berekend. De ICC is een maat voor de betrouwbaarheid van herhaalde metingen die varieert tussen 0 en 1. De hier gebruikte ICC (2,1), ook bekend als *ICCagreement*, is de verhouding van patiëntvariantie (in de teller van de breuk) tot (in de noemer van de breuk) patiëntvariantie plus variantie die toe te schrijven is aan de andere twee variantiebronnen: meetmomenten en meetfouten.[7,13,14] Om te weten welk GIPscoreverschil betrouwbaar groter is dan de standaardmeetfout is voor elke GIP-subschaal de standaardmeetfout (*SEMagreement*) berekend, als de vierkantswortel uit de som van de twee variantiebronnen: meetmomenten en meetfouten.[15] Door de standaardmeetfout (SEMagreement) te vermenigvuldigen met √2 en de z-waarde 1,96 verkrijgt men de *minimal detectable difference* (MDD95), ook wel bekend als *reliable change index*.[7] MDD95 betekent dat 95% van de patiënten die tussen twee metingen in werkelijkheid niet zijn veranderd, een toevallige scorefluctuatie kunnen laten zien van MDD95 of minder. Verschilscores groter dan MDD95 wijzen met een waarschijnlijkheid van 95% op een betrouwbaar verschil, groter dan de standaardmeetfout.

Bland-Altman plots

Om opvallend grote verschillen tussen herhaalde metingen zichtbaar te maken werd het verschil tussen de eerste en de tweede score afgezet tegen het gemiddelde van deze scores. Verschilscores buiten een bandbreedte van twee standaarddeviaties rond het gemiddelde verschil werden gedefinieerd als 'uitzonderlijk'. [16] Deze analyse werd toegepast bij alle achttien observatieschalen (veertien GIP-subschalen en de vier afgeleide schalen voor hulpbehoevendheid, apathie, cognitie en affect). De Bland-Altman plots werden gebruikt om deelnemers te identificeren voor wie de premisse van 'stabiliteit' in de testhertestperiode zeer onwaarschijnlijk mocht worden geacht.

Deelnemers

In de onderzoeksperiode werden 109 achtereenvolgende deelnemers opgenomen in een hertestbestand. Hun leeftijd varieerde tussen 53 en 96 jaar (zie Tabel 1). Gemiddeld (mediaan) waren de deelnemende patiënten 15 weken in dagbehandeling, 50% tussen 11 en 22 weken.

Tabel 1 vermeldt ook gegevens uit de Cognitieve Screening Test (CST) en de Amsterdamse Dementie-Screeningstest (ADS). Beide methoden geven een indicatie van de ernst van de cognitieve stoornis van de deelnemende patiënten. De CST heeft twintig items en bevat oriëntatievragen, vragen naar bekende publieke personen (zoals de koningin, de ministerpresident) en vragen naar feiten uit voorbije jaren (zoals de jaartallen van de twee wereldoorlogen).[17] De ADS is een verzameling korte tests

voor onderzoek van episodisch geheugen, oriëntatie, uitvoerende mentale controle en visueel-constructieve vaardigheden. Tabel 1 geeft de gewogen ADS-scores, die kunnen variëren van -6 tot +4 (ADS3) of van -10 tot +8 (ADS5), waarbij hoge scores duiden op een relatief hoog niveau van cognitief functioneren.18 De meeste deelnemers hadden een dementie van het Alzheimertype (82%), 13% een vasculaire dementie, 3% dementie met een andere oorzaak, zoals ziekte van Parkinson. De duur van de dementie was gemiddeld 4,1 jaar (SD02,4). Bij 2% van de deelnemers was geen dementie vastgesteld, maar een affectieve stoornis, zoals depressie, of een enkelvoudige cognitieve stoornis.

Resultaten

Van de 109 uitgezette hertestformulieren werden er 57 (52%) ingevuld terugontvangen. Bij de 52 patiënten voor wie geen tweede GIP werd ingevuld waren er elf in de periode tussen test en hertest ernstig ziek geworden, bijvoorbeeld door een cerebrovasculair accident, of overleden. Negen patiënten waren om andere redenen gestopt met dagbehandeling. De overige 32 niet-ingevulde formulieren bleven langer dan twee maanden uit, vanwege vakantie, ziekte of tijdgebrek van de betrokken medewerkers.

De GIP-scores van de twee groepen zijn vergeleken (52 patiënten van wie geen, en 57 patiënten van wie wel een tweede GIP-formulier werd ingevuld). De mediaantoets liet voor twee van de achttien subschalen (Apathisch gedrag en cluster Apathie) een significant verschil zien (p e 0,05) in de richting van meer apathisch gedrag bij deelnemers voor wie geen hertestformulier beschikbaar was. De verschillen tussen de twee groepen varieerden in termen van de associatiecoëfficiënt 8 van 0,01 tot 0,19. Vanaf 800,30 is er sprake van een matig sterk verschil, 8 Q 0,50 wijst op een sterk verschil.[19,20] De conclusie is dat de scoreverdelingen van de twee groepen (met of zonder hertest) geen belangrijke verschillen laten zien.

Herhaalde metingen

Tussen de eerste en tweede invulling van de GIP verliepen gemiddeld (mediaan) 45 dagen. De helft van de hertestformulieren werd tussen 34 en 58 dagen ingevuld terugontvangen. Het kortste en langste interval was 26 resp. 95 dagen. In totaal waren er elf medewerkers die één tot tien ingevulde hertestformulierenretourneerden. Gemiddeld (mediaan) vulden zij vijf hertestformulieren in.

Tabel 1 **Deelnemerskenmerken (*N*=109, 57,8% vrouw)**

Variabele	Laagste	Q_1	Mediaan	Q_3	Hoogste
Leeftijd	53	75,5	80	84	96
Deelname aan dagbehandeling	6 weken	11 weken	15 weken	22 weken	> 5 jaar
CST[a)]	3,5	10,4	13,3	16	20
ADS_3[b)]	−5	−2	0	1	4
ADS_5[b)]	−5	−1	1	3	8

[a)] *N*=102; [b)] *N*=105

CST Cognitieve Screening Test; ADS_3 Amsterdamse Dementie-Screeningstest, bestaande uit de subtests Visueel Geheugen, Oriëntatie en Fluency. In de ADS_5 zijn hieraan nog de subtests Meander en Natekenen toegevoegd

Q_1=score behorend bij het 25-ste percentiel van de frequentieverdeling; Q_3=score behorend bij het 75-ste percentiel van de frequentieverdeling

Bij inspectie van de Bland-Altman plots werd één deelnemer gevonden die opvallend grote veranderingen tussen test- en hertestscores liet zien. Figuur 1 laat zien hoe deze deelnemer met een verbetering van vier punten op GIPsubschaal 4 (Decorumverlies) duidelijk afweek van de meeste andere patiënten.

Dezelfde patiënt toonde bij nog vier andere observatieschalen ‘extreme’ verschillen (meer dan 2SD boven of onder het gemiddelde verschil). Zijn gedrag werd bij de eerste meting veel minder gunstig beoordeeld dan bij de tweede meting, zeventig dagen later. Bij geen van de andere deelnemers kwamen dergelijke grote discrepanties op meer dan drie

observatieschalen voor. Daarom werd besloten deze ene patiënt voor de berekening van de testhertestbetrouwbaarheid buiten beschouwing te laten.

Deelnameduur

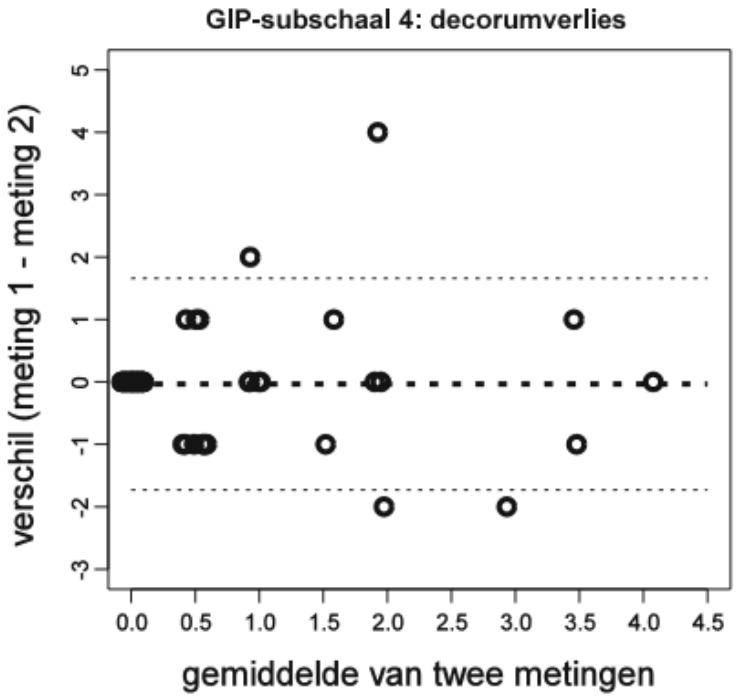
De 56 overige patiënten waren bij aanvang van het test-hertestonderzoek minimaal 59 dagen (bijna twee maanden) in dagbehandeling. Daarmee is de kans gering dat verschillen tussen herhaalde metingen met de GIP de gevolgen van gewenning of een eerste therapeutische invloed van dagbehandeling zouden weerspiegelen. In regressieanalyses met deelnameduur als onafhankelijke variabele en de testhertestverschilscores op de achttien GIP-schalen als afhankelijke variabelen, waren de regressiecoëfficiënten niet significant afwijkend van nul. Inspectie per puntenwolk liet gelijke verdelingen van verschilscores zien voor deelnemers die relatief lang, en voor hen die relatief kort in dagbehandeling waren.

Tijdsinterval tussen herhaalde metingen

Gelet op de variabele lengte van de testhertestperiode is ook de relatie tussen de duur van het interval tussen de herhaalde metingen en de grootte van de absolute verschilscores onderzocht. Getoetst werd de nulhypothese dat er geen verband zou zijn tussen de lengte van het interval en de grootte van de verschillen (positief of negatief) tussen herhaalde metingen. De nulhypothese kon voor slechts twee variabelen worden verworpen (p90,05; tweezijdig), en wel voor GIP-subschaal 9 (Zinloos herhalend gedrag) en de clusterscore Cognitie. Met een langer interval tussen de twee metingen werden de (absolute) verschillen voor Zinloos herhalend gedrag kleiner (r0 − 0,32; p00,017) en voor Cognitie groterr00,27; p00,047). Een correlatieanalyse met ruwe verschilscores liet voor geen van de schalen een systematisch verband tussen intervallengte en verschilscores zien.

Betrouwbaarheidscoëfficiënten (ICC) en standaardmeetfouten (SEM)

Tabel 2 laat zien dat de ICC varieert van 0,57 (voor GIP-subschaal 6: Incoherent gedrag) tot 0,93 (voor GIP-subschaal 14: Angstig gedrag). Tabel 2 bevat ook de standaardmeetfouten (SEM) voor de onderzochte GIP-schalen. De standaardmeetfout (SEM) is bij een betrouwbaarheidscoëfficiënt 9 0,75 kleiner dan de helft van een standaarddeviatie (SD). In dat geval neemt de nauwkeurigheid van een metingin verhouding tot die standaarddeviatie met meer dan 50% toe.6 Tabel 2 bevat twaalf schalen waarvoor de betrouwbaarheidscoëfficiënt (ICC) groter is dan 0,75. Het zijn vooral deze schalen die in de onderzochte steekproef een relatief stabiele meting van gedrag toelieten. De desbetreffende schalen zijn te vinden in het sociale gedragsdomein (voor drie van de vier GIP-schalen 1-4 werd een ICC 9 0,75 gevonden). In het cognitieve domein (GIP-schalen 5 tot 10) hadden twee schalen een ICC 9 0,75. In het affectieve domein (GIP-schalen 11–14) vonden we drie schalen met een ICC 9 0,75. De scores voor de clusters hulpbehoevendheid, apathie, cognitie en affect hadden alle een ICC van 0,82 of 0,83. Dit is een aanwijzing dat dergelijke clusterscores relatief stabiel en betrouwbaar gemeten kunnen worden.



Figuur 1 Verschillen tussen meting 1 en meting 2 op GIP-subschaal 4 (decorumverlies) in relatie tot het gemiddelde van de twee metingen. De stippellijnen zijn het gemiddelde van de verschilscores ± 2SD. *N*=57

Afgerond naar hele getallen variëren scoreverschillen (positief of negatief) die met een waarschijnlijkheid van 10% of minder aan toevallige fluctuaties kunnen worden toegeschreven van 1 voor GIP-subschaal 14 (Angstig) tot 4 voor GIP-subschaal 12 (Somber), Hulpbehoevendheid en Apathie (zie Tabel 2, kolom MDD90). Volgens een stringenter criterium (een verandering die uitstijgt boven een waarschijnlijkheid van 95% 'toevallige' fluctuaties) variëren de vereiste verschilscores (positief of negatief) van 1 voor GIP-subschaal 14 (Angstig) tot 5 voor Hulpbehoevendheid (zie Tabel 2, kolom MDD95).

Discussie

De primaire uitkomst van dit onderzoek is dat voor twaalf van de achttien observatieschalen van de GIP een betrouwbaarheidscoëfficiënt van minstens 0,75 werd gevonden. Met een betrouwbaarheidscoëfficiënt 9 0,75 neemt de precisie van een meting ten opzichte van de standaarddeviatie met meer dan 50% toe. Dit is een argument om een waarde van 0,75 als ondergrens voor een stabiele meting te hanteren.6 De observatieschalen die een relatief stabiele meting toelaten, werden gevonden in alle drie gedragsclusters waarover de GIP uitspraken toelaat: (sociale) activiteit, cognitie en beleving (affect). Voor de GIP-clusterscores Hulpbehoevendheid, Apathie, Cognitie en Affect, die items uit diverse schalen combineren, werden betrouwbaarheidscoëfficiënten van 0,82 of 0,83 gevonden, ruim boven de ondergrens van 0,75.

Tabel 2 GIP-scores (minimum en maximum in deze steekproef, gemiddelde M en standaarddeviatie SD) bij eerste (M_1; SD_1) en tweede invulling (M_2; SD_2). N=56

GIP-subschaal	Gedrag of stoornis	Min.-max.	M_1	SD_1	M_2	SD_2	ICC	SEM	MDD_{90}	MDD_{95}
1	Niet-sociaal	0–16	2,6	3,8	2,8	3,4	0,89	1,22	2,8	3,4
2	Apathisch	0–9	2,2	2,3	2,7	2,3	0,72	1,22	2,8	3,4
3	Bewustzijnsstoornis	0–14	1,0	2,5	1,4	3,1	0,88	0,98	2,3	2,7
4	Decorumverlies	0–4	0,5	1,0	0,6	1,1	0,79	0,48	1,1	1,3
5	Opstandig	0–6	0,7	1,5	1,0	1,8	0,73	0,88	2,0	2,4
6	Incoherent	0–5	1,1	1,3	1,2	1,4	0,57	0,90	2,1	2,5
7	Geheugenstoornis	0–10	3,6	2,7	3,9	2,5	0,80	1,16	2,7	3,2
8	Gedesoriënteerd	0–5	1,1	1,5	1,1	1,4	0,64	0,88	2,0	2,4
9	Zinloos herhalend	0–8	0,6	1,3	0,7	1,5	0,89	0,46	1,1	1,3
10	Rusteloos	0–6	1,9	1,6	1,8	1,7	0,68	0,94	2,2	2,6
11	Achterdochtig	0–11	0,9	2,2	1,1	2,6	0,91	0,73	1,7	2,0
12	Somber	0–10	2,1	2,6	3,0	3,0	0,76	1,40	3,3	3,9
13	Afhankelijk	0–7	1,9	1,9	2,2	1,9	0,63	1,17	2,7	3,2
14	Angstig	0–7	0,2	1,0	0,3	1,1	0,93	0,28	0,7	0,8
	Hulpbehoevendheid	3–17	6,4	3,9	6,6	3,8	0,82	1,63	3,8	4,5
	Apathie	0–15	2,8	3,5	3,0	3,3	0,83	1,38	3,2	3,8
	Cognitie	0–10	2,6	2,7	3,1	2,6	0,82	1,11	2,6	3,1
	Affect	0–14	1,8	2,4	2,6	2,9	0,83	1,11	2,6	3,1

Hoge (gemiddelde) scores wijzen op een hogere mate van expressie van genoemde gedragingen of stoornissen. De laatste vier kolommen geven de intraclass correlation coefficient (ICCagreement), de standaardmeetfout (SEMagreement), en de minimal detectable difference (MDD) berekend met een 90%- resp. 95%-waarschijnlijkheidsinterval (MDD_{90} en MDD_{95})

Over herhaalde metingen met de GIP bij deelnemers aan psychogeriatrische dagbehandeling is niet eerder gepubliceerd. Met behulp van de gegevens uit dit onderzoek kon voor elk van de gebruikte GIP-schalen worden vastgesteld welke scoreverandering nodig is voor een betrouwbaar verschil, dat uitstijgt boven verschillen vanwege niet-systematische meetfouten. De gegevens van dit onderzoek kunnen in de klinische praktijk worden gebruikt om in het individuele geval te beoordelen of er sprake is van een 'echte' verandering. Zie de twee casusvignetten voor toepassing in de klinische praktijk.

Beperkingen van het onderzoek

Dit onderzoek is uitgevoerd in een klinische praktijk waarin logistieke problemen een volgehouden en ononderbroken verzameling van betrouwbaarheidsgegevens in de weg stonden. Slechts voor iets meer dan de helft van de beoogde deelnemers werd een hertestformulier ingevuld. Uitval hing niet alleen met patiëntfactoren samen (ziekte, stoppen met dagbehandeling), maar ook met de extra administratieve last voor de betrokken medewerkers, waardoor het tijdig invullen van een hertestformulier niet altijd lukte.

De herhaalde invulling van de GIP werd na een interval van drie tot vier weken aangevraagd. In de praktijk verliep er meestal meer tijd tussen eerste en tweede invulling. Een vooraf gekozen duur van een interval tussen herhaalde metingen berust op

de afweging dat het interval lang genoeg moet zijn om oorspronkelijke waarnemingen mogelijk te maken, maar ook weer niet zo lang dat reële veranderingen in de toestand van de patiënt de resultaten gaan beïnvloeden. Sommige reële veranderingen kunnen zich zelfs binnen een korte periode voltrekken, zoals acute verwardheid (delier) of veranderingen in stemming, zoals depressie. Over een langer durende periode neemt de kans op reële veranderingen vanzelfsprekend toe, bijvoorbeeld door progressie van dementie. Ook zijn reële veranderingen mogelijk door de invloed van dagbehandeling zelf op gedrag, cognitie en stemming van de deelnemer of door gebeurtenissen waarvan de hulpverleners van de dagbehandeling geen weet hebben, zoals bijvoorbeeld verandering van medicatie door huisarts of medisch specialist.

In dit onderzoek zijn verschillende voorzorgsmaatregelen genomen om de invloed van reële veranderingen zo klein mogelijk te houden. Deelnemers waren minimaal zes weken in dagbehandeling, zodat de eerste effecten van deelname aan dagbehandeling al in de eerste meting zouden worden verdisconteerd. Bij evidente veranderingen in de gezondheidstoestand, ernstige gebeurtenissen thuis of tussentijdse afwezigheid wegens ziekte werd geen hertestformulier ingevuld. Bland-Altman plots werden gebruikt om deelnemers te detecteren die tussen de herhaalde metingen onwaarschijnlijk grote gedragsveranderingen lieten zien. Dit resulteerde in uitsluiting van één deelnemer. Ten slotte werden diverse correlatieanalyses uitgevoerd die lieten zien dat verschillen in deelnameduur en verschillen in de duur van het interval tussen herhaalde metingen geen systematische invloed hadden op de grootte van de verschilscores tussen beide metingen.

Hoewel niet voor alle GIP-schalen betrouwbaarheidscoëfficiënten 9 0,75 werden gevonden, is voor de diverse domeinen van sociaal gedrag, cognitie en stemming een keuze mogelijk voor schalen met een relatief hoge betrouwbaarheid, ook wanneer het gedrag betreft dat op korte termijn gemakkelijk fluctueert, zoals acute verwardheid (zie de GIP-schaal Zinloos herhalend gedrag) of stemming (zie de GIPsubschalen voor Achterdochtig, Somber en Angstig gedrag). Betrouwbaarheidscoëfficiënten G 0,75 maken de desbetreffende schaal niet per se onbruikbaar. Maar in verhouding tot de standaarddeviatie zijn er relatief grote scoreverschillen nodig om te kunnen concluderen tot een betrouwbaar verschil (zie Tabel 2).

Bij de hier vermelde betrouwbaarheidscoëfficiënten hoort de kanttekening dat zij afhankelijk zijn van de steekproef en de situatie waarin de coëfficiënten zijn bepaald.[21] De ingevulde hertestformulieren lieten een lichte oververtegenwoordiging zien van deelnemers met relatief lage GIP-scores. Aanvullend onderzoek is nodig bij deelnemers met hogere scores, die wijzen op moeilijk hanteerbaar gedrag of ernstige cognitieve stoornissen.

Conclusie

Het hier beschreven onderzoek ondersteunt de toepasbaarheid van de GIP voor de beoordeling van veranderingen in het gedrag van deelnemers aan psychogeriatrische dagbehandeling. Voor sommige (minder stabiele) subschalen zijn voor de detectie van een minimaal verschil relatief grote verschilscores nodig, maar dit nadeel wordt gecompenseerd door de hogere betrouwbaarheid van clusterscores in de vier domeinen hulpbehoevendheid, apathie, cognitie en affect.

Casusvignet 1

Anna (74) neemt ruim anderhalf jaar deel aan de dagbehandeling. Haar activiteitenbegeleidster maakt zich zorgen over veranderingen in haar stemming. Volgens GIP-subschaal 12 (Somber) zegt Anna zich nutteloos te voelen, is zij bang voor dingen die staan te gebeuren, en voelt zij zich regelmatig neerslachtig. De somscore 10 is drie punten hoger dan bij een eerdere invulling van de GIP, een half jaar geleden. Dit verschil valt nog binnen het 90%-predictie-interval. De score voor Affect, die bestaat uit een selectie van items uit de GIP-subschalen 11 (Achterdochtig), 12 (Somber), 13 (Afhankelijk) en 14 (Angstig) ismet vier punten toegenomen, van 4 naar 8. Het verschil is groter dan de minimal detectable difference (MDD95) met een 95%-predictie-interval in Tabel 2. De conclusie is dat Anna emotioneel minder goed in evenwicht is dan een half jaar geleden.

Casusvignet 2

Ben (75) heeft drie maanden nadat hij met dagbehandeling begon nog maar weinig contact met andere bezoekers. Dat komt tot uitdrukking in een score 8 op GIP-subschaal 1 (Niet-sociaal gedrag). Hij vindt dat het activiteitenprogramma te weinig tegemoetkomt aan zijn behoefte aan fitnesstraining. Gevraagd wat zijn deelname aan de dagbehandeling hem oplevert, zegt hij: ‘niks, ik heb er niet om gevraagd’. Hij zegt ook ‘geen idee te hebben hoe hij hier verzeild is geraakt’. Twee maanden later is hij met behulp van de fysiotherapeut aan fitnesstraining begonnen. Hij lijkt zich wat meer te hebben verzoend met zijn

deelname aan de dagbehandeling. Hij is vaker bereid anderen te helpen, begint uit zichzelf een gesprek, en luistert meer naar wat anderen vertellen. De score op Niet-sociaal gedrag is nu 4. Het verschil van vier punten is groter dan het 95%-predictie-interval (MDD95) in Tabel 2 zou voorspellen.

Auteurs

Han F. A. Diesfeldt

De Stichtse Hof, Vivium zorggroep, Laren, Netherlands
psycholoog, zelfstandig onderzoeker

Castricum

e-mail: h.diesfeldt@outlook.com

Literatuurlijst

1. Verstraten PFJ, Van Eekelen CWJM. In: Handleiding voor de GIP: Gedragsobservatieschaal voor de Intramurale Psychogeriatrie. Deventer: Van Loghum Slaterus; 1987.
2. De Jonghe JFM, Calis PJA, Boom-Poels PGM. Gedragsdimensies van oudere patiënten: Factorstructuur van de Gedragsobservatieschaal voor de Intramurale Psychogeriatrie (GIP). Tijdschrift voor Gerontologie en Geriatrie. 1996;27159-164.
3. De Jonghe JFM. Behavioral dimensions of dementia: Vrije Universiteit Amsterdam; 2001.
4. De Jonghe JFM, Kat MG, Rottier WPTJ, De Reus R. De Gedragsobservatieschaal voor de Intramurale Psychogeriatrie (GIP) en de klinische diagnose; een vergelijking met de BOP en NOSIE-30. Tijdschrift voor Gerontologie en Geriatrie. 1995;2624-29.
5. De Jonghe JFM, Kat MG, De Reus R. De validiteit van de Gedragsobservatieschaal voor de Intramurale Psychogeriatrie (GIP): een vergelijking met de BOP en NOSIE-30 in een psychiatrische observatiekliniek voor ouderen. Tijdschrift voor Gerontologie en Geriatrie. 1994;25110-116.
6. Streiner DL, Norman GR. Health measurement scales. A practical guide to their development and use. Fourth edition. Oxford: Oxford University Press, 2008.
7. Portney LG, Watkins MP. Foundations of clinical research. Applications to practice. Third edition. Upper Saddle River: Pearson Education, 2009.
8. Diesfeldt HFA. Over de interpretatie van veranderingen in BOP-scores bij longitudinaal onderzoek van individuele patiënten. Gerontologie. 1981;12212-217.
9. Shah A, Evans H, Parkash N. Evaluation of three aggression/agitation behaviour rating scales for use on an acute admission and assessment psychogeriatric ward. International Journal of Geriatric Psychiatry. 1998;13415-420. 10.1002/(SICI)1099-1166(199806)13:6<415::AID-GPS788>3.0.CO;2-A
10. Sommer OH, Engedal K. Reliability and validity of the Norwegian version of the Brief Agitation Rating Scale (BARS) in dementia. Aging and Mental Health. 2011;15252-258. 10.1080/13607863.2010.519318
11. Zuidema SU, Buursema AL, Gerritsen MGJM, Oosterwal KC, Smits MMM, Koopmans RTCM. Assessing neuropsychiatric symptoms in nursing home patients with dementia: reliable change index of the Neuropsychiatric Inventory and the Cohen-Mansfield Agitation Inventory. International Journal of Geriatric Psychiatry. 2011;26127-134. 10.1002/gps.2499
12. De Jonghe JFM, Ooms ME, Ribbe MW. Verkorte Gedragsobservatieschaal voor de Intramurale Psychogeriatrie (GIP-28). Tijdschrift voor Gerontologie en Geriatrie. 1997;28119-123.
13. Program to calculate intraclass correlation. http://www.stattools.net, 2012.
14. Intraclass correlation calculator. http://department.obg.cuhk.hk, 2012.
15. De Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. Journal of Clinical Epidemiology. 2006;591033-1039. 10.1016/j.jclinepi.2005.10.015
16. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet 1986(February 8):307–310.

17. De Graaf A, Deelman BG. In: Cognitieve Screening Test. Lisse: Swets en Zeitlinger; 1991.
18. Lindeboom J, Jonker C. In: Amsterdamse Dementie-Screeningstest. Lisse: Swets and Zeitlinger; 1989.
19. Cohen J. Statistical power analysis for the behavioral sciences. Second edition. Hillsdale: Lawrence Erlbaum, 1988.
20. Kline RB. Beyond significance testing. Washington DC: American Psychological Association; 2004.
21. Meyer P. Reliability. Oxford: Oxford University Press; 2010.