De N=1 statistiek achter het patient volg systeem in het FPC Dr. S. van Mesdag

Erwin Schuringa, Vera Heininga, Marinus Spreen

Inleiding

Het doel van een psychiatrische en psychologische behandeling is het stabi- liseren of verbeteren van gedrag en/of klachten. Om deze effecten te kunnen meten is in het FPC Dr. S. van Mesdag een routine outcome monitoring (ROM) systeem ontwikkeld (Schuringa, 2010). Dit wordt het patient volg systeem (PVS) genoemd. Het PVS bestaat uit een vragenlijst (Instrument voor Forensische Behandel Evaluatie: IFBE), die voor elke behandelbespre- king door een aantal beoordelaars wordt ingevuld en een speciaal voor het PVS ontwikkelde N=1 statistiek. Deze statistiek is opgezet om de scores van een patient op een bepaald tijdstip niet alleen te kunnen vergelijken met (de veranderingen ten opzichte van) een normgroep, maar ook met de individu- ele score op een eerder tijdstip. In dit laatste geval wordt de patient dan zijn eigen norm (Spreen, Stam & Bartels, 2003).
In dit artikel zal aan de hand van een voorbeelditem deze statistiek uitgelegd worden. Net als het artikel van Spreen et al. dient dit artikel voornamelijk gezien te worden als een uit een rij van methodologisch georienteerde studies naar de toepassing van zulke formele methoden voor het meten van effecten van therapie op patientniveau, ter ondersteuning van de behandeling (zie bijvoorbeeld ook Spreen, Timmerman, Ter Horst & Schuringa, 2010).

Methode

Voor de N=1 statistiek wordt gebruikgemaakt van een 17 punt Likert schaal (Likert, 1932). Deze schaal heeft vijf ankerpunten met beschrijvingen van gedrag dat bij die score hoort. Zo ontstaat er visueel een 5 puntsschaal met drie scoremogelijkheden tussen de ankerpunten in. Uit ervaring bleek namelijk dat er nog wel eens getwijfeld werd tussen twee scoremogelijkheden en dat er meestal drie twijfelplekken zijn: Als eerste tussen twee score-/anker- plekken in (het gedrag voldoet al niet meer aan de beschrijving behorende bij score 1, maar ook nog niet aan de beschrijving van score 2). De tweede twij- felplek is vlak boven een score-/ankerpunt (het gedrag is net iets beter dan de beschrijving van de score, maar weer niet goed genoeg voor een scorecatego- rie erboven). En als laatste vlak onder een score-/ankerpunt (het gedrag vol- doet nog niet helemaal aan de beschrijving van de score, maar zeker niet aan de beschrijving van een score lager). Zo krijg je dus tussen de scores 1 en 2 in
de scores 1 + , 1^ en 2-. De scorebalk waar antwoorden op gegeven kunnen worden ziet er als volgt uit:

Voor elke behandelbespreking vullen verschillende beoordelaars van FPC Dr. S. van Mesdag onafhankelijk van elkaar dezelfde vragenlijst over de patient in. Aan de hand van de scores volgt een rapportage die er voor het item probleeminzicht bijvoorbeeld als volgt uit ziet:

	Score Toen (0-4)	**Score NU (0-4)**	**Overeenstemming**	**Verandering Beoordelaars**
1 Probleeminzicht	2,00	**2,50**	hoog	▲ 3

De gemiddelde score van drie beoordelaars op meetmoment een was 2,00 en op meetmoment twee 2,50. De overeenstemming tussen de beoordelaars is hoog. En er is een betekenisvolle verbetering waargenomen in het probleeminzicht van de patient. Hieronder zal uitgelegd worden hoe deze gegevens tot stand komen.
Er waren drie beoordelaars op meetmoment een en ook op meetmoment twee. In onderstaande tabel staan de X-en voor de scores van de drie beoordelaars op meetmoment een en de O's voor de scores op meetmoment 2.

										O						
										O						
							X	**X**	**X**	**O**						
0	□	□	□	**1**	□	□	□	**2**	□	□	□	**3**	□	□	□	**4**

Met behulp van de N=1 statistiek worden verschillende maten uitgerekend: de gemiddelde score, de mate van overeenstemming en de mate van verandering.

De *gemiddelde score* is de scores van alle beoordelaars opgeteld en dan gedeeld door het aantal beoordelaars. Alle beoordelaars hebben hierin dus hetzelfde 'gewicht' ongeacht hun functie.

De mate van overeenstemming

Een van de aannames van de ontwikkelde N=1 statistiek is dat een beoorde- laar niet zeker is van zijn score. Hij zal altijd enigszins twijfelen of zijn score de werkelijkheid goed weergeeft. Een volgende aanname is dat hij per gegeven score niet verder zal twijfelen dan een scoremogelijkheid naar links (omlaag) en naar rechts (omhoog). Als hij wel zover had getwijfeld, dan had hij logischerwijs een andere score gegeven. Deze onzekerheid over de score wordt meegenomen in de N=1 statistiek. Zoals in tabel 1 te zien is krijgt elke beoordelaar er twee (twijfel)scores bij.

Tabel 1

In beeld ziet dit er zo uit:

Beoordelaar			Twijfel		Originele			Twijfel score +								
								B3								
							B2	B2	B2							
						B1	B1	B1	B3	B3						
0				1				2				3				4

Voor de mate van overeenstemming is een maat (de Jildou-index) ontwikkeld die ligt tussen de 0 en 1. Waarbij 0 maximale niet-overeenstemming is en 1 maximale overeenstemming uitdrukt. Hiervoor moet eerst de O-index (Overeenstemmings-index) uitgerekend worden. De O-index wordt vast- gesteld door te tellen hoe vaak een score voorkomt ten opzichte van het aan- tal scores (= aantal beoordelaars maal drie) in het kwadraat. In bovenstaande tabel komt de score 2 drie keer voor, 2- twee keer en 1^ een keer, et cetera. De O-index wordt dan:

			B3		
		B2	B2	B2	
	B1	B1	B1	B3	B3
Score	**1/**	**2-**	**2**	**2+-**	**2/**
Aantal	1	2	3	*2*	1
O-index	0.0123	0.0246	0.0370	1.0246	0.0123

$$O\text{ - }index \quad \frac{1 \qquad 2 \qquad 3 \qquad 2 \qquad 1}{(3x3)^2 + (3x3)^2 + (3x3)^2 + (3x3)^2 + (3x3)^2} = 0.111$$

De O-index is voor alle mogelijke combinrties van scores voor alle hoeveel- heden beoordelaars uitgerekend (tot maximaal 8 beooedelaars). 1-lier kveam een rangschikking uit voort van hoge O-index tot lage O-index per aantal beoordelaars. Echter, er waren ook groepen van scoremogelijkheden die dezelfde O-index hadden, maar qua overeenstemming visueel toch sterk van elkaar afweken. Onderstaande scoreverdelingen leveren bijvoorbeeld dezelfde O-index op, maar zijn qua overeenstemming toch echt verschillend:

B1	B1	B1	B2	B2	B2	B3	B3	B3								
0				1				2				3				4

B1	B1	B1					B2	B2	B2					B3	B3	B3
0				1				2				3				4

Om binnen deze groepen weer een rangschikking te maken is gebruikge- maakt van de standaarddeviatie. Dit is een maat die de individuele scores vergelijkt met het gemiddelde van de scores. Hoe groter de afstand tot het gemiddelde hoe hoger de standaarddeviatie. En dus ook hoe lager de over- eenstemming. De standaarddeviatie is veel gevoeliger voor extreme uitschie- ters dan de O-index. Daarom wordt er eerst op O-index gerangschikt en dan pas op standaarddeviatie. Door op deze manier de scores ook weer te rang- schikken binnen dezelfde O-index score kan er een totale rangschikking ge- maakt worden waarbij de hoogste overeenstemming het cijfer 1 krijgt en de laagste het cijfer 0 en alle overige een getal hier tussenin afhankelijk van de positie in de rangschikking. Dit getal is de Jildou-index.
De grens voor een goede overeenstemming is op 0.70 gesteld. Dit betekent dat de overeenstemming goed is als deze bij de beste 30% van alle mogelijke combinaties hoort. 0.70 is arbitrair gekozen, maar lijkt tot nu toe een werk- bare grens

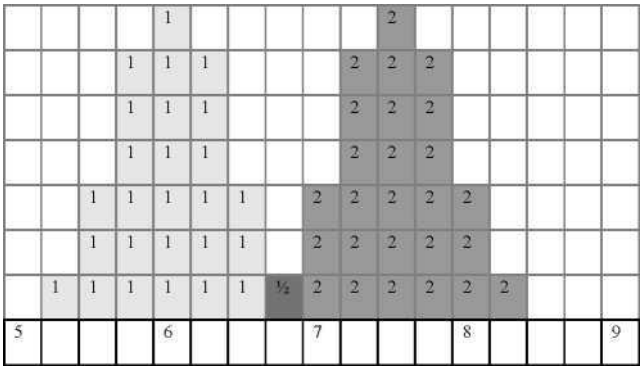
De mate van verandering

Voor de mate van verandering wordt ook gebruikgemaakt van de twijfel- scores. Het is dan mogelijk om van al de combinaties van deze scores (echte scores en twijfelscores) de som van de drie beoordelaars te berekenen (zie ta- bel 2).

Tabel 2

B1	B2	B3	som		B1	B2	B3	som		B1	B2	B3	som
1.5	2-	2	5.25		2-	2-	2	5.5		2	2-	2	5.75
1.5	2-	2+	5.5		2-	2-	2+	5.75		2	2-	2+	6
1.5	2-	2.5	5.75		2-	2-	2.5	6		2	2-	2.5	6.25
1.5	2	2	5.5		2-	2	2	5.75		2	2	2	6
1.5	2	2+	5.75		2-	2	2+	6		2	2	2+	6.25
1.5	2	2.5	6		2-	2	2.5	6.25		2	2	2.5	6.5
1.5	2+	2	5.75		2-	2+	2	6		2	2+	2	6.25
1.5	2+	2+	6		2-	2+	2+	6.25		2	2+	2+	6.5
1.5	2+	2.5	6.25		2-	2+	2.5	6.5		2	2+	2.5	6.75

Voor meetmoment twee kan hetzelfde worden gedaan. Je krijgt dan van 3 beoordelaars de volgende frequentiegrafiek:



Voor de berekening van de mate van verandering wordt de verhouding tus- sen niet-overlap scores en totaal aantal scores gerekend. Met andere woorden, hoeveel mogelijke scores op meetmoment twee komen overeen met scores op meetmoment een. Als de overlap perfect (100%) is betekent dit dat de scores op meetmoment twee gelijk zijn aan de scores op meetmoment een. Er heeft dan geen verandering plaatsgevonden. In bovenstaand geval zijn er 26 niet- overlap scores (een overlap bij score 6.75) van de 27 mogelijke overlapscores. Dit levert de volgende mate van verandering op: 26/27= 0.96. Wij hanteren een niet-overlap grens van 0.70 om te kunnen zeggen dat de patient beteke- nisvol is veranderd ten opzichte van het eerste meetmoment. Deze grens be- tekent dat 70% van de scores op meetmoment 2 niet voorkwamen op meet- moment 2. Deze grens is tot op heden arbitrair gekozen en zal nog verder onderzocht worden.

N=1 statistiek versus Reliable Change Index

Om te kijken of de N=1 statistiek en de gekozen grenzen bruikbaar zijn, is er een beperkte pilotstudy gedaan waarin bij 43 casussen de individuele verandering op basis van zowel de voor het PVS ontwikkelde N=1 statistiek als de Reliable Change Index (RCI; Jacobson andTruax, 1991) is vergeleken. Deze laatste is een index die in de forensische psychiatrie en de geestelijke gezond- heidszorg veel gebruikt wordt voor het meten van de individuele vorandering (zie bijvoorbeeld: Chakhssi, De Ruiter & Beensteif, 2010). Hierbij wordt de verandering van het individu vergeleken met de onbetrouwbiarheid van het instrument (de meetfout). Er is gebruikgemaakt van de eangepaste RCI zoals beschreven door Maassen, Bossemaen Brand (2009))

$$RCI = \frac{D - D_{,,}}{pi +) \; t \; i \; - \; \wedge)}$$

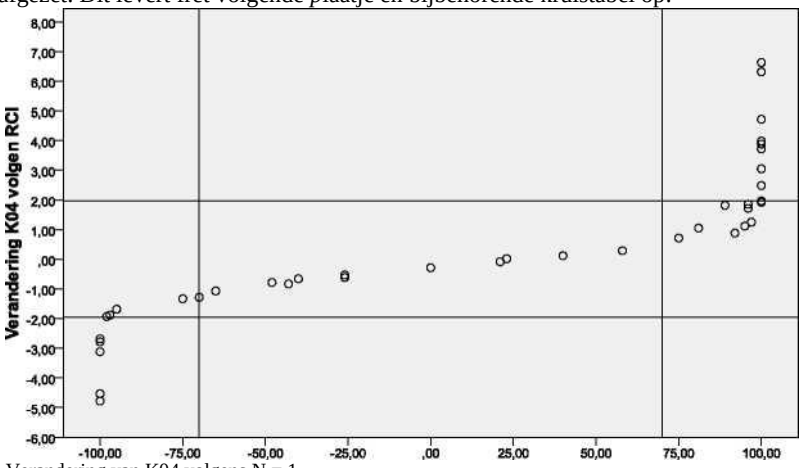
Waarbij *D.* de verschilscore van het individu is en D_c de gemiddelde verschil- score van de groep. S is de variantie van de groep op het eerste meetmoment en S^2 is de variantie op meetmoment twee. *r* is de correlatie tussen de beide

Y i xy

meetmomenten.

Een RCI boven de 1,96 (of onder de -1,96) wordt als een klinisch significan- te verandering gezien.

Voor het item K04 impulsiviteit zijn beide methodes uitgerekend en tegen elkaar afgezet. Dit levert fret volgende *p*laatje en bijbehorende kruistabel op:



Verandering van K04 volgens N = 1

		Verandering K04 volgens RCI			
		-1,00	,00	1,00	Total
Verandering K04 volgens N=1	-1,00	5	**4**	0	9
	,00	0	15	0	15
	1,00	0	**9**	10	19
Total		5	28	10	43

De S-vorm in de grafiek wordt veroorzaakt door de grenzen van tie mate ■van verandering. De maximale mate van overeenstemming is 100 en bij de RCI is deze oneindig. Deze gegevens laten zien dat er een lineair verband is tussen de N=1 statistiek en de RCI en dat de N=1 statistiek in 70% (30/43; gear- ceerde vakken) van de gevallen eenzelfde uitkomst heeft als de traditionele RCI. In 30% (13/43; vetgedrukte getallen) van de gevallen zeggen beide methodes iets anders. In het ideale geval zou een externe maat uitsluitsel kunnen geven over welke methode een juiste verandering aangeeft. Echter deze was tijdens de pilot niet voorhanden.

Discussie

De in dit artikel beschreven N=1 statistiek is een methode die nog in de kinderschoenen staat, daardoor zijn er aannames gedaan die nog geanalyseerd moeten worden, zoals de twijfelscores en de grenzen van mate van overeen-

stemming en verandering. Dat de N=1 statistiek redelijk vergelijkbaar is met de RCI is een zeer bemoedigende uitkomst, want in tegenstelling tot de RCI gaat men bij de N=1 statistiek niet uit van groepsgegevens. Deze zijn name- lijk vaak niet beschikbaar. De RCI is afhankelijk van de grootte en samenstel- ling van de gekozen groep. Een groep kan bestaan uit bijvoorbeeld alleen 'ge- zonde' personen, een mix van mannen en vrouwen, 'gewone' psychiatrische patienten, etc. De keuze van de groep zal de waarde van de RCI beinvloeden. Hierdoor wordt de patient niet volledig met zichzelf vergeleken maar met zichzelf ten opzichte van een groep.

Conclusie

De werkbaarheid van deze N=1 statistiek is in de praktijk tot nu toe zeer goed gebleken. De drie uitkomsten (gemiddelde score, mate van overeen- stemming en mate van verandering) zijn informatieve gegevens voor de eva- luatie van de behandeling van de individuele patient. Naast de vorm zoals weergegeven in tabel 1:

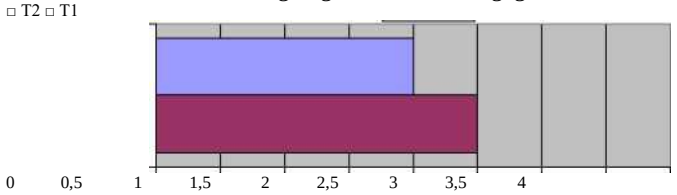
	Score **Score** (0-4) **NU**	**Overeenstemming***	**Verandering**	**Beoordelaars**
1 Probleeminzicht	2,00 **3,00**	hoog	▲	3

wordt in de terugkoppeling ook een tekstuele beschrijving van de gemiddel- de score gegeven:

Patient:

1 heeft enig probleeminzicht, handelt er niet altijd naar
dit was: heeft wel probleembesef, maar gedraagt zich niet hiernaar

En de scores worden ook nog in grafiekvorm weergegeven:



Dezelfde informatie wordt dus op drie manieren weergegeven. Dit is gedaan omdat verschillende behandelaars een verschillende voorkeur voor de weer- gave van de gegevens hebben. Voor een uitgebreide casusbeschrijving zie ar- tikel Schuringa (2011) in deze editie.
Naast dat de N=1 statistiek erg werkbare informatie oplevert voor behande- laars, blijkt dat de uitkomsten van de N=1 statistiek vergelijkbaar zijn met die van de RCI. De N=1 statistiek levert dus geen volledig andere informatie op dan de gangbare methode. Een groot pluspunt daarbij van de N=1 statistiek is dat deze geen groepsgegevens gebruikt en ook niet nodig heeft. In de praktijk kan het namelijk zo zijn dat deze groepsgegevens of niet relevant voor de patient zijn of zelfs helemaal niet voorhanden zijn.
In de nabije toekomst zal er echter nog meer onderzoek gedaan moeten worden

naar de plausibiliteit van de aannames, de gestelde grenzen en zal de eventuele meerwaarde ten opzichte van de RCI ook nog (verder) onderzocht moeten worden.

Literatuur

Chakhssi, F., de Ruiter, C. & Bernstein, D.(2010). Change during forensic treatment in psychopathic versus nonpsychopathic offenders. *Journal of Forensic Psychiatry & Psychology,* 21(5), 660-682.

Jacobson, N.S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology,* 59, 12-19.

Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology* 140, 1-55.

Maassen, G.H., Bossema, E., & Brand, N. (2009). Reliable change and practice effects: Outcomes of various indices compared. *Journal of clinical and experimental neuropsychology* 31(3), 339-352.

Schuringa, E. (2010). Routine Outcome Monitoring in het FPC Dr. S. van Mesdag. *GGzet Wetenschappelijk,* 14 (1), 27-35.

Schuringa, E. (2011). Een voorbeeldcasus uit het Patient Volg Systeem van het FPC dr. S. van Mesdag. *GGzet Wetenschappelijk.* (deze editie).

Schuringa, E., Bokern, H., Pieters, R. & Spreen, M. (2006). Atascadero Skills Profile Nederlandse Versie (ASP-NV). Een gedragsobservatie instrument voor de forensische psychiatrie. *GGzet Wetenschappelijk,* 10 (2), 40-46.

Spreen, M., Stam, G. & Bartels, A. (2003). N=”1” statistiek in de Dr. S. van Mesdagkliniek: Een praktisch voorbeeld van de SCL-90 Klachtenlijst als effect indicator van therapie. *GGzet Wetenschappelijk,* 7, (2), pp. 12-20.