Meetvariatie als bron van bias bij het benchmarken met verschillende ROM-instrumenten

M. BLANKERS, M. BARENDREGT, J.J.M. DEKKER

ACHTERGROND In de ggz worden momenteel behandeluitkomsten verzameld met verschillende meetinstrumenten. Dit heeft mogelijk implicaties voor de vergelijkbaarheid van deze uitkomsten tussen ggz-instellingen.

DOEL Bespreken van recent onderzoek naar de mate waarin de huidige (acht) gehanteerde meetinstrumenten leiden tot onderling vergelijkbare uitkomstmetingen van curatieve behandelingen voor volwassenen in de ggz.

METHODE Een combinatie van literatuurbespreking en empirisch onderzoek.

RESULTATEN De uitkomsten op basis van de acht meetinstrumenten bleken niet eenduidig te zijn: dezelfde cliënten leken met sommige instrumenten een sterkere klachtenafname te behalen dan met andere instrumenten.

CONCLUSIE De huidige benchmarkpraktijk binnen de ggz zou meer valide worden wanneer het aantal gehanteerde instrumenten wordt teruggebracht. Daarnaast is hoogwaardig kalibratieonderzoek nodig naar de vergelijkbaarheid van de resterende instrumenten. Idealiter gebruiken alle ggz-instellingen in de toekomst per zorgdomein hetzelfde instrument om hun uitkomsten vast te stellen.

TIJDSCHRIFT VOOR PSYCHIATRIE 58(2016)1, 55-60

TREFWOORDEN benchmarken, meetinstrumenten, routine outcome monitoring

ARTIKEL

In dit artikel bespreken wij op basis van bestaande literatuur en nieuw empirisch onderzoek in hoeverre behandeluitkomsten tussen ggz-zorgaanbieders zijn te vergelijken, wanneer deze zorgaanbieders gebruikmaken van verschillende meetinstrumenten in hun routine outcome monitoring(ROM)-proces. Het behoeft geen nadere uitleg dat het vergelijken van behandeluitkomsten die met verschillende instrumenten zijn gemeten niet ideaal is. Echter, dit is wel de praktijk die in de ggz is ontstaan de afgelopen jaren. De belangrijkste bezwaren tegen deze praktijk zijn eerder opgesomd, bijvoorbeeld door Van Os e.a. (2012) en Hafkenscheid en Van Os (2014).

Variatie in behandeluitkomsten

Verschillen in behandeluitkomsten tussen instellingen hangen samen met verschillen tussen instellingen in de middelen waarmee de behandeling wordt uitgevoerd (structuurkenmerken zoals personeelsbestand of budget) en met de wijze waarop de behandeling wordt uitgevoerd (proceskenmerken, bijvoorbeeld het gebruik van effectieve behandelprotocollen en protocoltrouw) (Donabedian 1968; Tansella & Thornicroft 1998; **FIGUUR 1**). Wanneer een instelling cliënten met ernstigere psychiatrische aandoeningen in behandeling heeft dan een andere instelling, kan dit cliëntkenmerk (ceteris paribus) ook tot verschillen in uitkomsten tussen deze instellingen leiden. Ten slotte kan de wijze waarop behandeluitkomsten worden

gemeten een bron van vertekening opleveren. Denk hierbij aan variatie in wijze en frequentie van meten, het responspercentage, of aan de gehanteerde *meetinstrumenten*. In dit artikel bespreken wij deze laatste potentiële bron van vertekening: meetkenmerken.

Kwaliteit van zorg vergelijken met benchmarkgegevens

Een belangrijke motivatie om benchmarkgegevens te verzamelen is om met de uitkomsten als geleide de kwaliteit van zorg te verbeteren (Barendregt 2015). Verschillen in behandeluitkomsten zijn echter alleen terug te voeren op verschillen in de kwaliteit van de zorg als ze samenhangen met proces- en/of structuurkenmerken. De variatie in gemeten behandeluitkomsten als gevolg van cliënt- en meetkenmerken dient daarom te worden geminimaliseerd.

In de huidige praktijk van de curatieve ggz voor volwassenen leveren instellingen de gegevens van één van acht verschillende instrumenten aan aan de Stichting Benchmark GGZ (SBG) met als doel dezelfde behandeluitkomst te meten. Op dit moment zijn dat: de *Brief Symptom Inventory* (BSI), de *Korte KlachtenLijst* (KKL), de *Symptom CheckList-90* (SCL-90), de *Depression, Anxiety and Stress Scale-21/-42* (DASS-21; DASS-42), de *Outcome Questionnaire-45-Symptomatische Distress* (OQ-45-SD), de *Hospital Anxiety and Depression Scale* (HADS), de *Symptom Questionnaire-48* (SQ-48) of de *Clinical Outcomes in Routine Evaluation-34-Problemen* (CORE-34-P).

De uitdaging is om uit de ruwe scores van deze verschillende meetinstrumenten vergelijkbare behandeluitkomsten te extraheren. In het kader van een pilot waarin zorgverzekeraar Achmea samen met SBG en een aantal ggz-zorgaanbieders (Arkin, GGZ Drenthe, HSK, Lentis, Max Ernst en Mentaal Beter) de zeggingskracht van behandeluitkomsten beproefde, is onder meer onderzoek gedaan naar in hoeverre dit mogelijk is.

VOORWAARDEN

Om de scores die zijn verkregen met verschillende meetinstrumenten te vergelijken moeten ze dezelfde meetpretentie hebben, in gelijke mate gevoelig zijn voor verandering en de scores moeten omgerekend kunnen worden naar een gemeenschappelijke meetschaal.

Meetpretentie

Voor vergelijkbare resultaten moeten instrumenten hetzelfde meten. Een eerste bron van onwenselijke verschillen kan ontstaan doordat scores op verschillende instrumenten onderling niet optimaal correleren. Correlaties tussen de verschillende instrumenten die SBG momenteel hanteert, schommelen rond de 0,80 (de Beurs e.a. 2012; Carlier e.a. 2012; Pijck e.a. 2014). Hoewel dergelijke correlaties als sterk kunnen worden beschouwd, duiden ze tegelijkertijd op onwenselijke variatie in de meetpretentie tussen de gehanteerde instrumenten.

Gevoeligheid voor verandering

Wanneer bij dezelfde cliënt het ene meetinstrument zou leiden tot een groter verschil in scores tussen begin- en vervolgmetingen (de zogenaamde 'delta's') dan het andere, dan is er sprake van een verschil in gevoeligheid voor verandering. In de praktijk is het nog niet haalbaar gebleken om voor alle instrumenten die nu worden aangeleverd aan de SBG in één studie een gelijkwaardige verandergevoeligheid aan te tonen. Wel zijn er enkele studies waarbij men op verschillende gegevensverzamelingen vergelijkingen tussen twee of meer instrumenten heeft uitgevoerd (zie **TABEL 1**).

Zo vergeleken De Beurs e.a. (2012) met separate gegevensverzamelingen de DASS-42 met de SCL-90, de BSI met de OQ-45 (tweemaal), de KKL met de OQ-45-SD en de CORE-34-P met de SCL-90. In slechts twee van de vijf gegevenssets leidde het gebruik van verschillende vragenlijsten tot vergelijkbare uitkomsten. Men concludeerde dat er niet zonder meer vanuit gegaan kan worden dat alle vergeleken meetinstrumenten gelijkwaardig verandergevoelig zijn.

Pijck e.a. (2014) hebben met andere meetgegevens de resultaten van De Beurs e.a. (2012) gerepliceerd voor de vergelijking tussen de KKL en de OQ-45-SD. Zij vonden dat de verandergevoeligheid van de KKL en de OQ-45-SD vergelijkbaar was. Carlier e.a. (2014) ten slotte rapporteren sterke correlaties tussen de verschilscores van de SQ-48, en de BSI

AUTEURS

MATTHIJS BLANKERS, (senior) wetenschappelijk medewerker, Arkin, Amsterdam, en Trimbos-instituut, Utrecht, en gastonderzoeker, afd. Psychiatrie, Academisch Medisch Centrum, Amsterdam.

MARKO BARENDREGT, theoretisch psycholoog en senior onderzoeker, Stichting Benchmark GGZ, Bilthoven.

JACK J. M. DEKKER, bijzonder hoogleraar klinische Psychologie, Vrije Universiteit, en hoofd afd. Onderzoek, Arkin, Amsterdam.

CORRESPONDENTIEADRES

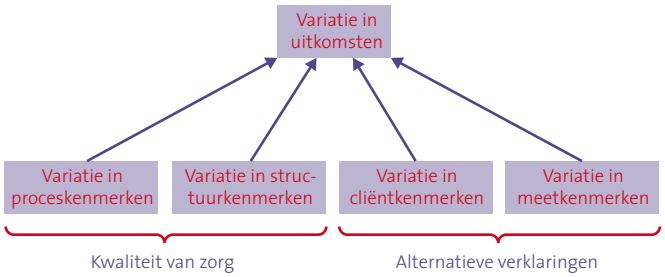
Dr. M. Blankers, Arkin, afd. Onderzoek, Postbus 75.848, 1070 AV Amsterdam.

E-mail: matthijs.blankers@arkin.nl

Geen strijdige belangen meegedeeld.

Het artikel werd voor publicatie geaccepteerd op 27-5-2015.

FIGUUR 1 **Bronnen van variatie in behandeluitkomsten**



en OQ-45-SD. Samengevat: er komt een niet-eenduidig beeld naar voren uit de verschillende studies naar de gelijkwaardigheid van de verandergevoeligheid van de voor benchmarking gebruikte instrumenten.

Gemeenschappelijke meetschaal

Ruwe scores op verschillende meetinstrumenten zijn op zichzelf niet onderling vergelijkbaar. Om voor benchmarking toch tot vergelijkbare scores te komen hanteert SBG de genormaliseerde T-score als instrumentvrije meetschaal. T-scores zijn in de jaren twintig van de vorige eeuw ontwikkeld door McCall (1922) en zijn een standaardiserende transformatie van ruwe scores, waarbij de resulterende score een gemiddelde waarde van 50 en een standaarddeviatie van 10 heeft.

Om de ruwe scores van een meetinstrument om te rekenen naar genormaliseerde T-scores definieert men voor elk meetinstrument een omrekenformule op basis van een kalibratiesteekproef die bestaat uit voormetingen van initiële DBC-trajecten uit de SBG-database. Om de vergelijkbaarheid optimaal te houden gebruikt men bij voorkeur gegevens van minimaal 1500 cliënten, behandeld bij ten minste vijf verschillende instellingen.

Op dit moment voldoen alleen de OQ-45-SD en de BSI aan deze strengste voorwaarden van de SBG. Voor de overige instrumenten zijn suboptimale kalibraties beschikbaar op basis van voormetingen verzameld binnen minder dan vijf instellingen. Voor bijvoorbeeld de relatief nieuwe SQ-48 is nog geen kalibratie beschikbaar omdat er nog onvoldoende metingen mee verricht zijn.

VERGELIJKBAARHEID DASS-21 EN CORE-34-P

In hoeverre zijn de (delta-)T-scores die zijn verkregen onder suboptimale kalibraties toch onderling vergelijkbaar? Om die vraag te beantwoorden en om de problemen te schetsen die kunnen ontstaan wanneer suboptimaal gekalibreerde instrumenten met elkaar worden vergeleken, deden wij bij ggz-instelling Arkin onderzoek hiernaar. Daarbij onderzochten wij de onderlinge vergelijkbaarheid van de scores op de DASS-21 en de CORE-34-P, twee instrumenten waarvan de T-scores suboptimaal zijn gekalibreerd op basis van twee verschillende steekproeven die verzameld zijn binnen één, respectievelijk twee instellingen. Om de onderzoeksvraag te beantwoorden werden in het uitgevoerde onderzoek beide instrumenten op twee tijdstippen afgenomen bij een steekproef van 175 cliënten die in behandeling waren bij Arkin.

TABEL 1 **Gepubliceerd onderzoek naar vergelijkbaarheid meetinstrumenten***

	BSI	CORE-34-P	DASS-21	DASS-42	KKL	OQ-45-SD	SCL-90	SQ-48
BSI								
CORE-34-P								
DASS-21		D						
DASS-42								
KKL								
OQ-45-SD					A,B			
SCL-90		A		A				
SQ-48	C					C		

*Gepubliceerd onderzoek: A = De Beurs e.a. (2012); B = Pijck e.a. (2014); C = Carlier e.a. (2012); D = Blankers e.a. (huidige artikel)

De ruwe scores op de DASS-21 en de CORE-34-P correleerden hoog ten tijde van de beginmeting (r = 0,87), wat steun geeft aan de vergelijkbaarheid van de meetpretentie van beide instrumenten. De effectgroottes (Cohens d) verkregen met beide instrumenten waren nagenoeg gelijk. Dit resultaat steunt de veronderstelling dat deze twee instrumenten gelijkwaardig verandergevoelig zijn.

Echter, na het omrekenen naar T-scores bleek dat er aanzienlijke verschillen optraden in de resultaten van de twee instrumenten, zoals te zien is in **TABEL 2**. Gevonden werd dat de berekende T-scores niet gelijkwaardig waren. De T-score van de CORE-34-P was in vergelijking met de T-score van de DASS-21 gemiddeld hoger en had een grotere standaarddeviatie. Dit effect trad op bij de begin- en de eindmetingen en ook bij de berekende verschilscores (de 'delta's'). De verschilscores weken significant van elkaar af ($t(175) = 3,48$; $p = 0,0006$), terwijl deze metingen steeds bij dezelfde cliënten en op hetzelfde moment werden afgenomen en dus idealiter hetzelfde resultaat zouden moeten laten zien. Wat de uitkomst van dit onderzoek duidelijk maakt, is dat DASS-21-T-scores niet dezelfde waarde hebben als CORE-34-P-T-scores, waardoor de resultaten die zijn vastgesteld met de DASS-21 en met de CORE-34-P flink van elkaar verschillen. De CORE-34-P laat immers bij dezelfde steekproef een grotere verandering in T-scores zien dan de DASS-21. De verklaring voor deze bevinding is dat de huidige T-scoreomrekenformules voor de CORE-34-P en de DASS-21 op steekproeven uit verschillende instellingen zijn gebaseerd. Het niet hanteren van een representatieve kalibratiesteekproef heeft klaarblijkelijk forse consequenties voor de vergelijkbaarheid van de gerapporteerde behandeluitkomsten.

DISCUSSIE

In dit artikel geven wij op basis van recent Nederlands onderzoek aan in welke mate bronnen van meetvariantie van invloed zijn op gemeten behandeluitkomsten – onafhankelijk van de kwaliteit van zorg. Variantie in meetpretentie, verandergevoeligheid en de kwaliteit van de kalibratie van instrumenten zijn van invloed op de onderlinge vergelijkbaarheid van ggz-instellingen zolang deze verschillende instrumenten hanteren om hun uitkomsten vast te stellen. Bij het evalueren van verschillen in gemeten behandeluitkomsten dient men rekening te houden met verschillen in meetkenmerken, zoals het gebruikte meetinstrument.

Op basis van de besproken bevindingen adviseren wij om nog strengere voorwaarden voor de representativiteit van de kalibratiesteekproef te hanteren. Immers, ook als aan de strengste voorwaarden voldaan is, valt niet uit te sluiten dat instrumenten toch niet optimaal vergelijkbaar zijn.

Beleidsmatig ontstaat een dilemma als we (om aan de praktijk van ROM in het veld tegemoet te komen) aan de ene kant instellingen willen vergelijken die verschillende instrumenten gebruiken en er aan de andere kant weinig ruimte mag bestaan voor twijfel aan de vergelijkbaarheid van de resultaten. Dit laatste is het geval omdat de uitkomsten in toenemende mate belangrijk worden voor (bijvoorbeeld) afspraken met zorgverzekeraars.

Een mogelijke oplossing die in grote lijn reeds eerder is voorgesteld (bijvoorbeeld door Van Os e.a. 2012), zou zijn om bij één (representatieve) steekproef van ggz-cliënten uit het zorgdomein 'volwassenen cure' alle te kalibreren meetinstrumenten gelijktijdig af te nemen. Op basis van deze steekproef worden vervolgens alle instrumenten gekalibreerd en worden de omrekenformules bepaald.

TABEL 2 **DASS-21 en CORE-34-P afgenomen bij dezelfde cliënten en onder gelijke meetcondities**

Meetmoment*	Instrument	N	Gemiddelde T-score	Standaard-deviatie	Cohens d (binnen cliënten) (50%-interval)
1e meting	DASS-21	175	50,1	10,2	-
	CORE-34-P	175	63,2	18,3	-
2e meting	DASS-21	175	46,7	9,7	-
	CORE-34-P	175	57,0	18,5	-
Verschil 1e-2e meting	DASS-21	175	3,4	8,4	0,42 (0,37-0,46)
	CORE-34-P	175	6,1	15,1	0,41 (0,36-0,46)

*Het tijdsinterval tussen de 1e en 2e meting bedroeg gemiddeld 138 dagen (SD 62).

Om er zeker van te zijn dat verschilscores van de instrumenten onderling ook vergelijkbaar zijn, moet men de afname van alle instrumenten na een vastgestelde periode herhalen. Zo kan men vaststellen dat de verschilscores verkregen met de verschillende instrumenten inderdaad gelijk(waardig) zijn.

Daarnaast is het wenselijk om het aantal voor benchmarking ondersteunde instrumenten af te bouwen van acht tot slechts enkele. Hoe meer instrumenten men gebruikt, hoe groter de kans dat hiertussen (aanzienlijke) variatie bestaat in meetpretentie of -gevoeligheid. Wanneer de resterende ondersteunde instrumenten na kalibratie aantoonbaar een gelijke meetpretentie, meetgevoeligheid en meetschaal hebben, is een bron van bias weggenomen uit het benchmarkproces.

Vanuit het perspectief van de meetintegriteit zouden idealiter alle ggz-instellingen binnen een zorgdomein scores op hetzelfde instrument moeten aanleveren aan de SBG ten behoeve van de landelijke benchmark. Dit zal echter op praktische bezwaren van de professionals stuiten, omdat zij graag meer vrijheid in het ROM-proces krijgen. Ook bij het ontwikkelen van nieuwe meetdomeinen, zoals de door het veld gewenste toevoeging van stoornisspecifieke uitkomstmaten, spelen deze uitdagingen een rol. Ook hier zou men voor een zuivere benchmark idealiter een keuze moeten maken voor één instrument per stoornis, terwijl in de dagelijkse praktijk verschillende instrumenten gebruikt worden. Gezien de toenemende nadruk op kwaliteit in de ggz, onder meer verwoord in het recente zorgplan van het ministerie van VWS (Schippers 2015), en gezien de in dit artikel besproken beperkingen aan het meten met verschillende instrumenten, hopen wij dat de discussie over het beperken tot één benchmarkinstrument per meetdomein snel heropend kan worden.

LITERATUUR

- Barendregt M. Benchmarken en andere functies van ROM: back to basics. Tijdschr Psychiatr 2015; 57: 517-25.
- Beurs E de, Barendregt M, Flens G, van Dijk R, Huijbrechts I, Meerding WJ. Vooruitgang in de behandeling meten – Een vergelijking van vragenlijsten voor zelfrapportage. MGV 2012; 67: 259-65.
- Carlier IVE, Meuldijk D, van Vliet IM, van Fenema EM, van der Wee NJA, Zitman FG. Empirische evidence voor de effectiviteit van routine outcome monitoring; een literatuuronderzoek. Tijdschr Psychiatr 2012; 54: 121-8.
- Donabedian A. The evaluation of medical care programs. Bull N Y Acad Med 1968; 44: 117-24.
- Hafkenscheid A, van Os J. Naar een deugdelijke ROM. MGV 2014; 69: 20-8.
- McCall WA. How to measure in education. New York: Macmillan; 1922.
- 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.
- Pijck L, Deen M, van den Berg J, Huijbrechts I, Korrelboom K. De veranderingsgevoeligheid van OQ-45 en KKL bij ROM. MGV 2014; 69: 31-6.
- Schippers EI. Kamerbrief over verbeteren kwaliteit en betaalbaarheid zorg. Den Haag: Ministerie van Volksgezondheid, Welzijn en Sport; 2015. https://archive.today/GsAsX.
- Tansella, M, Thornicroft, G. A conceptual framework for mental health services;the matrix model. Psychol Med 1998; 28: 503-8.

SUMMARY

Benchmarking using different measurement instruments and the management of measurement variability

M. BLANKERS, M. BARENDREGT, J.J.M. DEKKER

BACKGROUND In mental health care centres in the Netherlands outcome data are collected using a variety of outcome instruments. This may have implications for the comparability of outcome results between different centres.

AIM To discuss recent findings regarding the extent to which the eight instruments currently used in clinical practice report comparable results.

METHOD Our study is based on a combination of literature review and empirical research.

RESULTS The results obtained with the eight instruments are not equivalent. Patients symptom reductions appear larger with some instruments than with others.

CONCLUSION The current practice of benchmarking in the Dutch mental health system would have greater validity if the number of different instruments would be reduced. State-of-the-art calibration studies are necessary to validate the comparability of the remaining instruments. Ideally, all mental health centres will soon use one instrument per care domain to measure treatment outcome.

TIJDSCHRIFT VOOR PSYCHIATRIE 58(2016)1, 55-60

KEY WORDS benchmarking, measurement instruments, routine outcome monitoring