ROM, benchmark en casemix in een vrije praktijk

R.M. MOSTERMAN

ACHTERGROND De benchmarkmethodes van routine outcome monitoring-trajecten (ROM-trajecten) door de Stichting Benchmark GGZ (SBG) zijn niet getoetst in onafhankelijk onderzoek. Over benchmarkonderzoek bij vrijgevestigden is niet gepubliceerd. De Alliantie Kwaliteit in de GGZ (AkwaGGZ) heeft de doorontwikkeling van ROM als kwaliteitsinstrument op zich genomen: een uitgelezen kans voor verbetering.

DOEL Replicatie van (delen van) de SBG-studies in een vrijgevestigde solopraktijk.

METHODE Observationeel onderzoek van herhaalde metingen met de Symptom Check List-90 (SCL-90).

RESULTATEN Gedurende ruim 12 jaar vulden 644 patiënten bijna 1200 maal de SCL-90 in. De ROM-groep (≥ 2 metingen, n = 339) verschilde van de niet-ROM-groep (≤ 1 meting, n = 210) op een aantal patiënt- en behandelkenmerken, maar niet op de voormeting. Betrouwbare verbetering deed zich voor bij 73% van de ROM-patiënten; klinisch herstel bij 45%. Om te kunnen benchmarken, werd de ROM-groep in drie subgroepen verdeeld, die drie zogenaamde zorgaanbieders representeerden. Benchmark toonde aan dat deze zorgaanbieders verschilden in behandelverloop en -resultaat. De nameting werd voorspeld door voormeting, comorbiditeit en een lage of gemiddelde opleiding. Hiermee werd 33% van de variantie verklaard. Casemixcorrectie leidde tot grotere verschillen tussen de zorgaanbieders.

CONCLUSIE Toepassing van ROM in de klinische praktijk is, op patiëntniveau, een nuttig hulpmiddel om het behandelproces te sturen en terminering te bepalen. Binnen-behandelaarbenchmark van geaggregeerde metingen maakt patronen en valkuilen zichtbaar en is leerzamer dan tussen-behandelaarsbenchmark. De SBG-onderzoeken bevatten methodologische tekortkomingen. Casemixvariabelen zijn niet generaliseerbaar over populaties; correctie daarop kan beter achterwege worden gelaten.

TIJDSCHRIFT VOOR PSYCHIATRIE 62(2020)1, 27-36

TREFWOORDEN benchmark, casemixcorrectie, ROM, vrijgevestigden

ARTIKEL

De Alliantie Kwaliteit in de GGZ (AkwaGGZ) heeft de doorontwikkeling van routine outcome monitoring (ROM) in de ggz op zich genomen. Vanaf 2011 verwerkte, bewerkte en toetste de Stichting Benchmark GGZ (SBG) vragenlijst-, persoons- en behandelgegevens van honderdduizenden ggz-patiënten. Het verschil tussen eerste en laatste meting werd als behandeluitkomst en kwaliteitsmaat gedefinieerd. Het doel van deze grootschalige gegevensverzameling was '*het verbeteren van de zorg in de ggz door transparantie te bieden over de behandelresultaten*' (website SBG).

Barendregt (2015) stelt dat benchmarken moet worden onderscheiden van de andere functies van ROM, zoals klinisch beleid, maatschappelijke verantwoording, wetenschappelijk onderzoek, financiering van zorg op basis van uitkomsten en keuze-informatie voor patiënten. Elk van deze activiteiten vraagt een andere wijze van validering en terugkoppeling van de gegevens. De vraag is of dat onderscheid in de praktijk ook daadwerkelijk gemaakt wordt. De wettelijk verplicht gestelde aanlevering van ROM-gegevens, aanvankelijk alleen voor ggz-instellingen, stuitte al

vanaf het begin op vele bezwaren en is vanaf 2017 vrijwel geheel stil komen te liggen. Vooral over het gebruik van ROM als benchmark en over de schending van de privacy van patiënten ontstonden hevige discussies. De petitie Stop Benchmark met ROM werd bijna 7000 keer getekend. Ook de Algemene Rekenkamer (2017) betoonde zich zeer kritisch en achtte ROM niet geschikt als basis voor bekostiging.

Nu de SBG is ontbonden en Akwaggz aan zet is, lijkt het tijd om lering te trekken uit de ervaringen die tot dusverre met het verzamelen en analyseren van ROM-gegevens is opgedaan. Eén verandering is al aangekondigd: patiënten dienen expliciet toestemming te verlenen voor het aanleveren van hun gegevens aan de Akwaggz. Hiermee ontstaat een nieuw probleem in de vorm van een potentiële selectiebias: in hoeverre verschillen privacy-bezwaarde patiënten van hen die geen bezwaar hebben?

Het is nog onduidelijk in hoeverre Akwaggz tegemoet zal komen aan de andere ernstige en goed onderbouwde bezwaren tegen gebruikmaking van ROM als benchmark, zoals methodologische en medisch-ethische bezwaren (o.a. Hafkenscheid & Van Os 2013, 2016, 2018; Van Os e.a. 2012, 2017). Er zijn slechts weinig onderzoeken op basis van Nederlandse ROM-uitkomsten gepubliceerd. Van controleerbare wetenschap is geen sprake: data zijn niet openbaar of toegankelijk gemaakt voor onafhankelijke onderzoekers.

ROM in eigen praktijk?

Sinds in januari 2017 ook voor vrijgevestigden aanlevering van ROM-gegevens verplicht werd, heb ik nieuwe patiënten in mijn praktijk hiervoor schriftelijk toestemming gevraagd. Alle patiënten tekenden hiertegen bezwaar aan. Hierdoor heeft de SBG geen ROM-gegevens van mij ontvangen.

Aangezien ik geïnteresseerd ben in kennisvergroting door wetenschappelijk onderzoek, het nut inzie van maatschappelijke verantwoording en keuzemogelijkheden voor patiënten, en ook nieuwsgierig ben naar de precieze manier waarop met ROM wordt gebenchmarkt, heb ik de ROM-gegevens uit mijn praktijk geanalyseerd. Het betreft een observationeel onderzoek met herhaalde klachtmetingen, waarbij de methodes waar mogelijk zijn gerepliceerd op basis van de volgende drie studies:

- Een observationeel onderzoek door De Beurs e.a. (2015), waarin zij uitkomstindicatoren (Delta-T, effectgrootte, statistisch betrouwbare verandering, klinisch significante verandering, percentages van verbetering en herstel) vergeleken. Auteurs concludeerden onder meer dat Delta-T en Cohens effectgrootte nagenoeg overeenkwamen.
- Een observationeel onderzoek door De Beurs e.a. (2018), waarin zij behandeluitkomsten bij patiënten met angst- en depressieve stoornissen vergeleken. Wanneer proceskenmerken als behandelduur en behandelkosten werden meegewogen, leidde dit tot grotere verschillen tussen zorgaanbieders.
- Een observationeel onderzoek door Warmerdam e.a. (2017) met een dataset van meer dan 31.000 patiënten ter bepaling van en correctie voor casemixvariabelen. Op basis van regressieanalyses bleek het klachtenniveau bij aanvang van de behandeling de belangrijkste voorspeller van behandeluitkomst te zijn, gevolgd door een beperktere invloed van GAF-score, leeftijd, sociaaleconomische status en enkele diagnoses.

De gegevensanalyses in mijn onderzoek zijn gericht op drie kernthema's die het ROM-debat kenmerken: ROM als behandelondersteuning; benchmark met ROM; identificatie van casemixvariabelen en toepassing van casemixcorrectie.

METHODE

Patiënten

In de periode oktober 2006-december 2018 werden 810 personen naar mijn praktijk verwezen. Bij 166 (20,5%) van hen kwam het niet tot een zorgcontact. De gegevens van de 644 patiënten vormden het onderzoeksmateriaal. De patiëntkenmerken worden vermeld in **TABEL 1**.

Gebruikt instrument

De *Symptom Check List-90* (SCL-90; Arrindell & Ettema 1986, 2005) is een multidimensionele zelfbeoordelingsschaal voor de acht belangrijkste klachtengebieden in de psychiatrische symptomen van volwassenen. Voor het onderhavige onderzoek werd, in navolging van de werkwijze van de SBG, alleen de totaalscore gebruikt, die het algehele niveau van psychisch/lichamelijk disfunctioneren weergeeft. De betrouwbaarheid (Cronbachs alfa) bij de voormeting was 0,97.

AUTEUR

INEKE MOSTERMAN, klinisch psycholoog-psychotherapeut, Psychologenpraktijk Elf, Zwolle.

CORRESPONDENTIEADRES

R.M. Mosterman, Psychologenpraktijk Elf, Bloemendalstraat 5, 8011 PJ Zwolle.
E-mail: i.mosterman@psychologenpraktijkelf.nl

Geen strijdige belangen meegedeeld.

Het artikel werd voor publicatie geaccepteerd op 17-7-2019.

TABEL 1 **Cliëntkenmerken**

Cliëntkenmerk	Allen (n = 644)	ROM compleet[a] (n = 339)
Geslacht (vrouw)	72%	76%
Gem. leeftijd (SD), in j	40 (13)	39 (13)
Partnersituatie (gehuwd/samenwonend)	60%	56%
Opleiding		
basis/vmbo	11%	7%
havo/vwo/mbo	38%	38%
hbo	40%	45%
wo	11%	11%
Werksituatie		
werknemer/zelfstandige	57%	56%
werkloos	7%	7%
zw/wao/wia	16%	18%
studerend	10%	11%
niet werkzaam/pensioen	9%	8%
DSM-as I		
stemming	38%	42%
angst	33%	35%
relatie	8%	4%
aanpassing	6%	7%
somatoform	5%	5%
overige	11%	8%
DSM-as II[b] (persoonlijkheidsproblemen)	44%	41%
DSM-as III[c] (somatiek)	29%	28%
DSM-as IV (complicerende factoren)	79%	78%
Gem. DSM-as V-score (SD)		
Begin behandeling	60 (7)	59 (6)
Einde behandeling	70 (10)	74 (9)
Comorbiditeit (met as I, II, en/of III)	58%	55%
Beëindiging		
Regulier	70%	91%
Verwijzing	7%	4%
Uitgevallen	19%	5%
Duur behandeling		
Gem. aantal weken (SD)	33 (33)	50 (35)
Gem. aantal sessies (SD)	14 (13)	20 (15)
Testafnames		
0	11%	0%
1	33%	0%
≥2	53%	100%

[a]Uitgezonderd 25 cliënten van wie de behandeling nog niet werd afgerond.

[b]Zowel persoonlijkheidsstoornis als trekken daarvan.

[c]Zowel enkelvoudige als complexe somatiek.

Procedure

Vragenlijsten, waaronder standaard de SCL-90, werden na het eerste gesprek door de patiënt ingevuld (n = 574). Het behandelplan werd gebaseerd op hulpvraag, anamnese, biografie en testresultaten. Het evalueren en hertesten volgde het natuurlijk beloop. Gemiddeld genomen vonden vervolgmetingen plaats na een half jaar (n = 361), een jaar (n = 142), anderhalf jaar (n = 67), twee jaar (n = 35), en twee jaar en drie maanden (n = 20).

Doorgaans werd een behandeling afgesloten wanneer op basis van onder andere de SCL-90 kon worden vastgesteld dat er voldoende en stabiele verbetering was bereikt. Beëindiging vond ook plaats wanneer er geen sprake meer was van enige progressie, waarbij het risico op terugval werd meegewogen. De laatste SCL-90-afname vond plaats vóór het einde van de therapie, zodat er nog gelegenheid was om zorgvuldig af te ronden.

Om te kunnen benchmarken, werd ervoor gekozen drie subgroepen te onderscheiden op basis van de perioden waarin zich veranderingen voordeden in organisatie (eerstelijnszorg/basis-ggz versus psychotherapie/specialistische ggz) en financiering (uurtarieven versus DBC-financiering) van de zorg, alsmede de ervaringsjaren en vervolgkwalificaties van de behandelaar. Deze subgroepen werden betiteld als drie verschillende zorgaanbieders:

- zorgaanbieder A, GZ-psycholoog-eerstelijnspsycholoog, ruim 12 jaar ervaring, was actief in de periode 2006-2009, en richtte zich voornamelijk op kortdurende behandelingen voor 281 patiënten met doorgaans lichtere problematiek, die volgens uurtarief werden gedeclareerd;
- zorgaanbieder B, > 16 jaar ervaring, werkte in de periode 2010-2013 als klinisch psycholoog-psychotherapeut met 228 patiënten conform de DBC-systematiek;
- zorgaanbieder C, eveneens klinisch psycholoog-psychotherapeut, > 20 jaar ervaring, rondde 110 behandelingen af in de periode 2014-2018, voornamelijk in de specialistische ggz.

(Voor alle duidelijkheid: alle patiënten werden feitelijk door mij behandeld.)

Statistische analyses

De analyses werden uitgevoerd met SPSS. Voor analyses over het onderdeel ROM werd de ruwe totaalscore van de SCL-90 gebruikt.

Voor het onderdeel benchmark werden de ruwe begin- en eindscores (RS) omgezet in genormaliseerde T-scores op basis waarvan Delta-T werd berekend, conform De Beurs (2010) en de formule uit de Factsheet Meetinstrumenten SCL-90 van de SBG (2017): [-22,079927576494+(0,6958209486682 35*RS)+(-0,0021314633335502*RS*RS)+(0,0000026376143166 7505*RS*RS*RS)].

In tegenstelling tot De Beurs e.a. (2015), die RCI en CS samenvoegden tot vijf elkaar overlappende categorieën, hetgeen bepaling van percentages bemoeilijkt, werden RCI en CS afzonderlijk gedefinieerd: RCI verbeterd werd $\Delta T > 5$; onveranderd werd ($\Delta T \leq 5$ en $\Delta T \geq -5$); verslechterd werd $\Delta T \leq -5$; CS (hersteld) werd (voormeting > 42,5 en nameting ≤ 42,5 en $\Delta T > 5$). Cohens effectgroottes werden berekend volgens de formule: $d = M_{diff}/([SD_1+SD_2]/2)$.

Voor het onderdeel casemixcorrectie werden met categorale (dummy-)variabelen en continue variabelen enkelvoudige en meervoudige regressieanalyses (methode enter) toegepast, conform Warmerdam e.a. (2017). Met de uiteindelijke regressieformule werd de casemixcorrectie toegepast.

Verschillen tussen zorgaanbieders of tussen subgroepen patiënten werden met enkelvoudige variantieanalyses getoetst.

RESULTATEN

Van de 644 patiënten werd bij 70 geen SCL-90 afgenomen. Redenen hiervoor waren bijvoorbeeld vroege uitval, verwijzing, onvoldoende beheersing van de Nederlandse taal of laaggeletterdheid.

Van nog eens 210 patiënten was er wel een voormeting, maar geen nameting. Uiteindelijk resulteerde dit in 339 volledige ROM-trajecten met twee of meer afnames van de SCL-90, uitgezonderd de 25 patiënten die eind 2018 nog in behandeling waren.

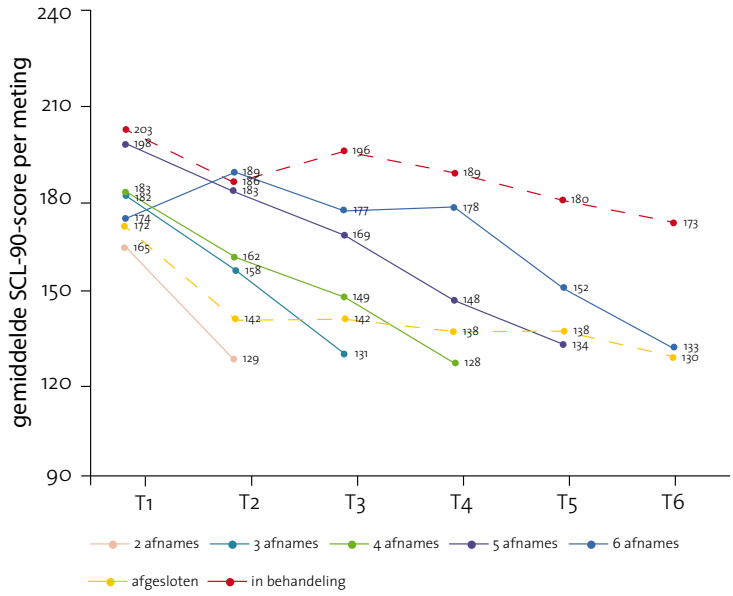
ROM als behandelondersteuning

Teneinde het klachtenverloop gedurende de therapie te kunnen volgen, werden de patiënten verdeeld in vijf groepen aan de hand van het aantal afnames ≥ 2 tijdens hun behandeling (ononderbroken lijnen in **FIGUUR 1**). Het klachtenverloop (voor-, tussen- en nametingen) is weergegeven in **FIGUUR 1**.

De groepen verschilden significant op de voormeting ($F(4) = 4{,}114$; $p < 0{,}01$). Uit post-hocanalyses (volgens Bonferroni) bleek dat patiënten met meer dan twee metingen een significant hogere beginscore hadden dan degenen die na de tweede meting waren gestopt. Bij de nameting werden geen significante verschillen tussen de groepen gevonden. In **FIGUUR 1** zijn tevens alle afgesloten ROM-trajecten (n = 339) opgenomen plus de metingen van patiënten die nog in behandeling waren (n = 25). (N.B. De zesde meting (M = 173) van deze 25 patiënten betrof dus nog niet de eindscore.)

Patiënten met een hogere beginscore bleken een langere behandeling nodig te hebben. De duur tussen eerste en laatste sessie van de afgesloten behandelingen verschilde significant over de groepen, zowel in aantal sessies ($F(4) = 122{,}428$; $p < 0{,}001$), variërend van M = 13 (SD 8) bij twee

FIGUUR 1 **Klachtenverloop (gemiddelde SCL-90-score per meting)**



metingen tot M = 64 (SD 16) bij zes metingen, als in aantal weken ($F(4) = 95{,}362$; $p < 0{,}001$); M = 33 (SD 22) tot M = 127 (SD 21). Bonferronivergelijkingen toonden aan dat dit voor alle groepen gold, met uitzondering van de vijfde en zesde afname, waarbij het verschil in duur niet significant was.

Het gemiddelde behandeleffect was groot (d = 1,12). Bij 73% van de patiënten was er sprake van een statistisch betrouwbare verbetering, van wie 45% als klinisch hersteld aangemerkt kon worden, 25% veranderde onvoldoende en bij 2% was sprake van een verslechtering.

Hoewel het merendeel van de patiënten een geleidelijke klachtvermindering liet zien, bleek er een subgroep van patiënten te zijn die een fluctuerend klachtenverloop vertoonde (beginscore < tussenscore, n = 63 in de ROM-groep). Deze patiënten hadden een beduidend slechtere prognose. Bij hen was, in vergelijking met de andere groepen, sprake van een significant lagere beginscore ($t(337) = -2{,}827$; $p < 0{,}01$), bij een relatief hogere eindscore ($t(72{,}80) = 4{,}154$; $p < 0{,}001$) en een kleinere verschilscore tussen voor- en nameting ($t(337) = 8{,}117$; $p < 0{,}001$) (in fluctuerende groep $M_{verschil} = 8{,}81$ (SD 40,08); in ROM-groep $M_{verschil} = 42{,}42$ (SD 36,69). Ook had de groep met flucuerend beloop een langere behandelduur ($t(81{,}83) = -5{,}189$; $p < 0{,}001$ (weken) en $t(72{,}30) = 5{,}189$; $p < 0{,}001$ (sessies)).

ROM en benchmark

De genormaliseerde T-scores zijn per zorgaanbieder vermeld in **TABEL 2**.

Over de periodes (= per zorgaanbieder) nam de ernst van de aanmeldklachten toe, zoals bleek uit toenemend hogere beginscores, maar het verschil tussen de zorgaanbieders was niet significant. De eindscores verschilden wel significant ($F(2{,}336) = 3{,}397$; $p < 0{,}05$). De verschillen op ΔT waren echter niet significant. De rangorde van het behandelresultaat van de zorgaanbieders was B-A-C.

ROM en casemixcorrectie

Voorafgaand aan het bepalen van de invloed van de casemix werd eerst nagegaan in hoeverre de patiënten die geen volledig ROM-traject hadden doorlopen (n = 280), verschilden van degenen van wie wel uitkomstmaten beschikbaar waren (n = 339). Het bleek dat het ontbreken van ROM-gegevens significant vaker voorkwam bij mannen ($\chi^2(1) = 6{,}677$; $p < 0{,}05$); bij alleenstaanden ($\chi^2(1) = 8{,}241$; $p < 0{,}01$); en bij laag opgeleiden ($\chi^2(3) = 16{,}219$; $p < 0{,}01$); minder vaak bij patiënten die betaald werk hadden of arbeidsongeschikt waren, maar vaker bij patiënten die om andere redenen niet werkten ($\chi^2(4) = 11{,}471$; $p < 0{,}05$). Bij patiënten met relatieproblemen werd minder vaak ROM toegepast ($\chi^2(14) = 46{,}602$; $p < 0{,}001$).

Er werden geen significante verschillen gevonden op de DSM-IV-assen 2, 3 of 4. De niet-ROM-groep had bij aanvang een hogere GAF-score dan de ROM-groep ($t(536{,}301) = 3{,}051$; $p < 0{,}01$), terwijl de GAF-score bij afsluiting relatief lager was ($t(617) = -9{,}744$; $p < 0{,}001$).

De behandelduur van de niet-ROM-groep was beduidend korter, zowel in het aantal sessies ($t(388{,}777) = -16{,}852$; $p < 0{,}001$), als in het aantal weken dat de behandeling duurde ($t(442{,}780) = -17{,}847$; $p < 0{,}001$). Deels had dit te maken met het verhoudingsgewijs hoger aantal patiënten die uitvielen en het hogere aantal doorverwijzingen ($\chi^2(2) = 135{,}166$; $p < 0{,}001$).

TABEL 2 **Benchmark met genormaliseerde T-scores**

Zorgaanbieder	N	Voormeting M (SD)	Nameting M (SD)	ΔT M (SD)	d	r	% RCI +	o	-	% CS
Zonder casemixcorrectie										
A	110	45,58 (7,53)	36,36 (7,58)	9,22 (7,68)	1,22	0,48	70	27	3	46
B	157	46,88 (9,02)	36,72 (7,22)	10,16 (8,47)	1,25	0,42	69	29	2	46
C	72	48,07 (7,67)	39,14 (7,98)	8,93 (7,17)	1,14	0,58	82	15	3	40
Alle	339	46,71 (8,31)	37,11 (7,55)	9,60 (7,96)	1,21	0,49	73	25	2	45
Met casemixcorrectie										
A	110	45,58 (7,53)	36,27 (7,56)	9,30 (7,68)	1,23	0,49	72	25	3	46
B	157	46,88 (9,02)	36,50 (7,17)	10,38 (8,46)	1,28	0,48	71	27	2	48
C	72	48,07 (7,67)	39,93 (8,14)	8,14 (7,26)	1,03	0,58	74	21	6	32
Alle	339	46,71 (8,31)	37,16 (7,63)	9,55 (7,99)	1,20	0,50	72	25	3	44

Zorgaanbieder A: periode 2006-2009, B: periode 2010-2013; C: periode 2014-2018.

Van 75% (n = 210) van de niet-ROM-groep was wel een voormeting aanwezig. Het bleek dat de beginscore significant lager was dan die van de ROM-groep ($t(394{,}308) = -1{,}664$; $p < 0{,}05$).

Vervolgens werd voor de ROM-groep onderzocht welke patiëntvariabelen van invloed waren op de nameting, conform de procedure van Warmerdam e.a. 2017. Hiertoe werden patiëntvariabelen uit **TABEL 1** afzonderlijk getoetst in enkelvoudige lineaire regressieanalyses. Een hogere voormeting, een lage of gemiddelde opleiding, arbeidsongeschiktheid, as 2- en as 3-problematiek en comorbiditeit bleken elk afzonderlijk hogere eindscores te voorspellen. Een hoge opleiding, werkzaam zijn en een hogere GAF-score voorspelden significant lagere eindscores (zie **TABEL 3**, kolom 2).

Deze significante variabelen werden vervolgens in een meervoudig regressiemodel geanalyseerd (zie **TABEL 3**, kolom 3). Er bleven uiteindelijk vier casemixvariabelen over: voormeting, opleiding laag, opleiding middel en comorbiditeit (zie **TABEL 3**, kolom 4). De toegevoegde waarde van de voormeting was 25%; met de overige drie variabelen tezamen werd dit 33% ($F(3{,}334) = 42{,}822$; $p < 0{,}001$). De regressiecoëfficiënten zijn vermeld in **TABEL 3**.

Met de regressieformule [14,018 + 0,433∗(gestandaardiseerde beginscore) + 5,838∗(opleiding laag) + 1,631∗(opleiding middel) + 3,348∗(comorbiditeit)] werd de verwachte eindscore berekend. De *relative performance factor* (RPF), de ratio tussen werkelijke en verwachte eindscore, werd voor elke zorgaanbieder bepaald. Bij een RPF<1 is de prestatie beter dan verwacht. Zorgaanbieders A, B en C hadden een RPF van respectievelijk 0,998, 0,994 en 1,020. Per zorgaanbieder werd de werkelijke eindscore vermenigvuldigd met hun RPF, wat de gecorrigeerde eindscore opleverde.

Door de casemixcorrectie was de gemiddelde eindscore van de ROM-groep nagenoeg gelijk gebleven (de correlatie ongecorrigeerde en gecorrigeerde nameting was r = 0,998). De verschillen tussen de zorgaanbieders op de nameting waren echter iets groter geworden: ongecorrigeerd $F(2{,}336) = 3{,}397$; $p < 0{,}05$ versus gecorrigeerd $F(2{,}336) = 6{,}268$; $p < 0{,}01$. De gecorrigeerde eindscores en afgeleide maten zijn vermeld in **TABEL 2**.

ROM en proceskenmerken

Conform De Beurs e.a. (2018) werd nagegaan welke invloed de duur van de behandeling had op de uitkomst per periode. Voor het berekenen van de ratio duur/behandeluitkomst werden duur en ΔT gestandaardiseerd. Een ratio < 1 betekent dat per eenheid tijd er meer eenheid behandeluitkomst (ΔT) bereikt wordt, ofwel meer resultaat in minder tijd.

De resultaten zijn weergegeven in **TABEL 4**. Te zien is dat er bij verbeterde of herstelde patiënten sprake was van een relatief snellere klachtvermindering. De zorg van zorgaanbieder A vroeg gemiddeld genomen de minste tijd per eenheid verbetering en van zorgaanbieder C de meeste. De verschillen waren echter niet significant.

Ten slotte werden de behandelresultaten onderzocht bij enkele subgroepen. De ernst van de klachten (drie groepen met respectievelijk beginscore < 45; ≥ 45 en ≤ 55; en > 55) bleek significant te verschillen op zowel voormeting, nameting als ΔT (respectievelijk $F(2{,}336) = 743{,}438$; $p < 0{,}001$, $F(2{,}336) = 38{,}984$; $p < 0{,}001$ en $F(2{,}336) = 63{,}963$; $p < 0{,}001$). De primaire as I-diagnoses (stemming, angst en overige) verschilden significant op de voormeting ($F(2{,}336) = 11{,}669$; $p < 0{,}001$) en de ΔT ($F(2{,}336) = 7{,}033$; $p < 0{,}010$), maar niet op de nameting.

TABEL 3 **Voorspellers van de nameting: casemixvariabelen en regressiecoëfficiënten**

Variabele	enkelvoudige regressie B (SE)	meervoudige regressie B (SE)	meervoudige regressie uiteindelijke casemix B (SE)
Intercept			14,02
Voormeting	0,46*** (0,43)	0,41*** (0,05)	0,43*** (0,04)
Sekse	0,14 (0,96)		
Leeftijd	-0,02 (0,03)		
Partnersituatie	-0,13 (0,83)		
Opleiding laag	4,44** (1,62)	5,75*** (1,39)	5,84*** (1,37)
Opleiding middel	2,03* (0,84)	1,57* (0,73)	1,63* (0,71)
Opleiding hoog	-3,06*** (0,81)		
Werkzaam	-2,08* (0,82)	0,44 (0,83)	
Arbeidsongeschikt	2,29* (1,06)	-0,06 (1,07)	
Overige niet werkzaam	0,89 (0,94)		
As 1 stemming	0,70 (0,83)		
As 1 angst	0,69 (0,89)		
As 1 overig	-1,59 (0,92)		
As 2 persoonlijkheidsproblemen	3,54*** (0,81)	0,71 (0,93)	
As 3 somatiek	0,92 (0,91)		
As 4 psychosociale problemen	2,46* (0,98)	-0,05 (0,87)	
As 5 GAF	-0,41*** (0,06)	-0,04 (0,07)	
Comorbiditeit	3,98*** (0,80)	2,69** (0,93)	3,35*** (0,68)

***p < 0,001;

**p < 0,01;

*p < 0,05

DISCUSSIE

In de klinische praktijk is ROM niet alleen het op patiëntniveau passief monitoren van effecten, maar tevens een hulpmiddel om actief het behandelverloop te beïnvloeden en het moment van terminatie te bepalen, mede door feedback aan de patiënt. De onderzoeksresultaten tonen aan dat ROM zich daarnaast op geaggregeerd niveau uitstekend leent voor wetenschappelijk onderzoek naar onderliggende tendensen en processen. Dit stemt overeen met de bevindingen uit de Leiden Routine Outcome Monitoring Studie (Noorden e.a. 2014, Schat e.a. 2016) en de studie van De Jong (2016) met data van vrijgevestigden.

Een belangrijke bevinding is de wijze waarop het klachtenniveau zich manifesteert: bij een geleidelijk afnemend klachtniveau is de prognose veel beter dan bij een fluctuerend klachtniveau. Ook is duidelijk geworden dat er – gemiddeld en op het eerste gezicht – nauwelijks winst te behalen valt na de tweede meting.

Wanneer we echter naar subgroepen patiënten en naar het specifieke verloop van de klachten kijken, komt een genuanceerder beeld naar voren. Voor sommige patiënten loont het de moeite om door te behandelen: bij hen toont een verbetering zich pas na de vierde meting.

Voorts is gebleken dat in de onderzochte populatie de meeste klachtenreductie te zien is bij ernstiger klachten en bij stemmingsstoornissen, hoewel dit langer kan duren. Patiënten met een lagere beginscore eindigen het snelst op een eindscore die binnen de niet-klinische range valt.

Bij eenzelfde instrument maakt het in principe niet zoveel uit of ruwe scores dan wel genormaliseerde T-scores gebruikt worden. Standaardisering maakt het mogelijk om verschilscores (Delta-T) tussen zorgaanbieders te vergelijken, juist en idealiter bij gebruik van verschillende meetinstrumenten.

Dit laatste bleek echter niet mogelijk: vragenlijsten blijken te verschillen in de mate van veranderingssensitiviteit (Blankers e.a. 2016). De keuze van de SBG om het aantal te gebruiken instrumenten drastisch te beperken heeft de motivatie om te participeren vermoedelijk verminderd, vooral bij hen die al jarenlange expertise hadden opge-

TABEL 4 **Duur per uitkomst (gecorrigeerde ΔT), per verbeterde patiënt en per herstelde patiënt**

	duur		per ΔT		per verbeterde patiënt			per herstelde patiënt	
	N	M (SD)	M (SD)	ratio[a]	N	M (SD)	ratio[a]	N	M (SD)	ratio[a]
Sessies per zorgaanbieder										
A	110	17,26 (13,23)	1,21 (10,39)	1,02	79	16,86 (12,82)	0,90	51	17,88 (13,37)	0,87
B	157	20,55 (15,40)	2,12 (8,21	1,03	111	20,21 (14,73)	0,91	74	20,78 (14,71)	0,88
C	72	22,64 (13,90)	8,33 (54,04)	1,12	53	24,23 (14,82)	1,03	23	27,43 (14,78)	0,98
Alle	339	19,93 (14,51)	3,14 (26,21)	1,05	243	20,00 (14,35)	0,93	148	20,82 (14,52)	0,89
Weken per zorgaanbieder										
A	110	44,44 (34,00)	2,86 (30,61)	1,02	79	42,99 (33,04)	0,90	51	45,24 (34,87)	0,88
B	157	51,31 (35,96)	4,98 (17,33)	1,03	111	51,46 (35,45)	0,92	74	52,57 (34,48)	0,88
C	72	55,11 (33,52)	18,96 (119,81)	1,11	53	59,23 (35,01)	1,02	23	66,96 (34,44)	0,98
Alle	339	49,88 (34,96)	7,26 (59,11)	1,04	243	50,40 (34,96)	0,93	148	52,28 (35,10)	0,89

Zorgaanbieders A: 2006-2009; B: 2010-2013; C: 2014-2018.

[a] Gestandaardiseerde duur gedeeld door de gestandaardiseerde gecorrigeerde uitkomst.

bouwd met instrumenten die, zoals de SCL-90, uit de minimale dataset zijn verwijderd.

Benchmark is vooral leerzaam en nuttig als men vergelijkt met zichzelf, de onderliggende data kan bestuderen en op basis daarvan verbeteringen nastreeft. Wanneer geen zicht geboden wordt op de kenmerken en processen achter gemiddelden (die door de SBG niet werden teruggekoppeld), wordt men niets wijzer van het zich blindstaren op hogere of lagere Delta-T van anderen of op rangorde in zorgaanbieders.

Beperkingen

De SBG-studies vertoonden enkele methodologische tekortkomingen. In het onderzoek van De Beurs e.a. (2015) heb ik de berekeningen van RCI en CS moeten aanpassen teneinde overlap uit te sluiten.

In het onderzoek van De Beurs e.a. (2018) was de operationalisering van de behandelduur weinigzeggend: deze werd berekend in weken (inclusief de wachttijd tussen intake en behandeling), zonder specificatie van aantal sessies per week, hetgeen een substantieel verschil kan maken. In relatie tot de kosten (afhankelijk van het perspectief: de opbrengsten) zou feitelijk conform de DBC-systematiek per minuten directe tijd gerekend moeten worden om een realistisch beeld te krijgen.

In het onderzoek door Warmerdam e.a. (2017) gaan de nauwkeurige en stapsgewijze berekeningen van de casemixcorrectie voorbij aan het klassieke statistische gegeven dat regressieanalyses op samengestelde databestanden, zoals van verschillende zorgaanbieders, tot een vertekening leiden als gevolg van de grootte van de verschillende bestanden. De in aantal cases grootste bestanden wegen het zwaarst in de regressiecoëfficiënten. Dit betekent dat de zorgaanbieders die de meeste data aanleverden, feitelijk de norm bepaalden.

Bovendien wijzigt het gehele regressiemodel met iedere toegevoegde variabele, zoals ook blijkt uit de onderzoeken van Noorden e.a. (2014), Schat e.a. (2016) en De Jong (2016), die elk weer andere voorspellers vonden. In het huidige onderzoek waren het voormeting, comorbiditeit, een laag en gemiddeld opleidingsniveau die de uitkomsten beïnvloedden. Kortom, met uitzondering van de voormeting, zijn de meeste voorspellers van behandelresultaat niet generaliseerbaar, en zouden niet mogen leiden tot algemeen toegepaste correcties.

Elke bewerking van ruwe data brengt ruis met zich mee en gaat ten koste van transparantie. Door het toepassen van casemixcorrecties is niet meer duidelijk hoever de aangepaste resultaten afstaan van de oorspronkelijke resultaten. Er is dan geen sprake van absolute gegevens, maar van contextafhankelijke data. Zorgaanbieder C behandelde, zoals te zien is in **TABEL 2**, patiënten met een relatief hoge beginscore. Ook het aantal patiënten met comorbiditeit was verhoudingsgewijs hoog. Voor deze beide casemixvariabelen is gecorrigeerd, waardoor Delta-T lager is geworden en er dus minder verbetering en herstel was. Dit is de, tegenintuïtieve, prijs die kleinere, afwijkende zorgaanbieders betalen voor casemixcorrectie.

De betrouwbaarheid van Delta-T, die volgens De Beurs e.a. (2018) over het algemeen uitstekend is, waarmee zij eerdere bevindingen van Blankers (2016) negeren, is mede afhankelijk van de betrouwbaarheid van het gebruikte instrument. Dat neemt niet weg dat de methode van zelfrapportage per definitie subjectief is en geen verwijzing naar de objectieve toestand is, maar veeleer de representatie van de beleving ervan (Hafkenscheid & van Os 2018). Er kan sprake zijn van responstendenties die tot vertekening en minder betrouwbare gegevens leiden.

Benchmarken en behandelaar

Het in de benchmark negeren van op zijn minst een behandelaarsperspectief bij het vaststellen van de uitkomst van de behandeling is uitermate merkwaardig: '*Patiënt en clinicus zijn het namelijk niet per definitie eens over de aard en ernst van de aanmeldingsproblemen of over de effectiviteit van behandeling.*' (Hafkenscheid & Van Os 2013).

Daarmee lijkt de validiteit van Delta-T in het geding. Delta-T is in feite niets meer of minder dan de mate van klachtverandering tijdens een behandeling. Deze te bestempelen als *het* behandelresultaat in plaats van als *een van de* mogelijke behandeluitkomsten, is een te grote stap. Laat staan de nog grotere stap om deze kwantitatieve symptoomreductie dan ook meteen maar als *de* kwaliteit van de behandeling te betitelen. Indien hetzelfde instrument waarmee feedback wordt verschaft zowel als monitorinstrument (predictor) wordt gebruikt en als effectmaat (criterium), is contaminatie van deze onafhankelijke en afhankelijke variabelen voor de hand liggend (Hafkenscheid & Van Os 2018).

Geestelijke gezondheidszorg speelt zich af op het uiterst complexe en vaak smalle raakvlak tussen personen die zich niet laten standaardiseren. De relatie tussen behandelaar en patiënt is uniek en wordt gevormd door karakteristieken van beiden: de interactie van persoonlijkheid, ervaring, deskundigheid, aandoening, waarbij de een de ander zorg aanbiedt en de zorgontvanger zowel subject als object van deze zorg is. De patiënt zal zich hoe dan ook met de behandeling bemoeien, waardoor er sprake is van wederzijdse beïnvloeding. Mensen zijn geen kopieerapparaten (Barendregt 2015), waarbij de winst afhankelijk is van gestandaardiseerde productieprocessen en van materialen die niet tegensputteren. Benchmarken is kwaliteitsmanagement van organisaties en processen, niet van behandelaars en behandeluitkomsten.

CONCLUSIE

ROM als behandelondersteunend instrument is van grote waarde in de klinische praktijk. Het inzetten van ROM als benchmark heeft deze kwaliteit bijna de das omgedaan. Het heeft, zoals Delespaul en Cnubben (2017) terecht stellen, ROM veranderd van '*een intrinsiek gemotiveerd bottom-up proces voor de patiënt in je behandelkamer en de patiënten uit de eigen caseload, tot een top-down opgelegde administratieve verplichting.*'

Akwaggz zegt in te zetten op kwaliteit en op het ontwikkelen van kwaliteitsindicatoren die daadwerkelijk nuttig zijn voor professional én patiënt. Helaas blijft ook deze organisatie vasthouden aan de elkaar bijtende doelen van ROM, die voor zoveel onduidelijkheid, verwarring en teleurstelling hebben geleid: ondersteuning in de behandeling, intercollegiaal leren, vergelijkingsinformatie voor patiënten en voor zorginkoop. Dat kan en moet beter.

Data en syntaxen kunnen worden opgevraagd bij de auteur.

LITERATUUR

- Algemene Rekenkamer. Bekostiging van de curatieve geestelijke gezondheidszorg. Den Haag: Algemene Rekenkamer; 2017.
- Arrindell WA, Ettema JHM. SCL-90. Handleiding bij een multidimensionele psychopathologie-indicator. Amsterdam: Pearson; 2005.
- Barendregt M. Benchmarken en andere functies van ROM: back to basics. Tijdschr Psychiatr 2017; 57: 517-25.
- Blankers M, Barendregt M, Dekker JJM. Meetvariatie als bron van bias bij het benchmarken met verschillende ROM -instrumenten. Tijdschr Psychiatr 2016; 58: 55-60.
- Beurs E de. De genormaliseerde T-score. Een 'euro' voor testuitslagen. MGV 2010; 65: 684-95.
- Beurs E de, Barendregt M, Rogmans B, Robbers S, van Geffen M, van Aggelen-Gerrits M, Houben H. Denoting treatment outcome in child and adolescent psychiatry: a comparison of continuous and categorical outcomes. Eur Child Adolesc Psychiatry 2015; 24: 553-63.
- Beurs E de, Warmerdam EH, Oudejans SCC, Spits M, Dingemanse P, de Graaf, SDd, e.a. Treatment outcome, duration, ands costs: a comparison of performance indicators using data from eight mental health care providers in The Netherlands. Adm Policy Ment Health 2018; 45: 212-23.
- Delespaul P, Cnubben W. Benchmark ROM: de verkeerde soort administratie. De Psycholoog 2017; 52; 49-52.
- Hafkenscheid A, van Os J. Huidige ROM doet afbreuk aan valide kwaliteitsmeting. Tijdschr Psychiatr 2013; 55: 179-81.

- Hafkenscheid A, van Os J. Wat ieder die betrokken is bij ROM zich over de metingen moet realiseren. Tijdschr Psychiatr 2016; 58: 388-96.
- Hafkenscheid A, van Os J. Twee misvattingen over ROM. De Psycholoog 2018; 53: 34-44.
- Jong K de. Behandelresultaten van vrijgevestigde psychologen en psychotherapeuten. Tijdschr Psychother 2016; 42: 308-17.
- Noorden, MS van, Giltay EJ, van der Wee, NJA, Zitman, FG. De Leiden Routine Outcome Monitoring Studie: beloop van stemmings-, angst- en somatoforme stoornissen in de poliklinische praktijk. Tijdschr Psychiatr 2014; 56: 22-31.
- Os J van, Kahn R, Denys D, Schoevers RA, Beekman, ATF, Hoogendijk WJG, e.a. ROM: gedragsnorm of dwangmaatregel? Tijdschr Psychiatr 2012; 54: 245-53.
- Os J van, Berkelaar J, Hafkenscheid A, e.a. Benchmarken: doodlopende weg onder het mom van 'ROM'. Tijdschr Psychiatr 2017; 59: 247-50.
- Schat A, van Noorden M. Voorspellers van behandelresultaat in angststoornissen: ROM in de praktijk. PsyXpert 2016; 2: 4: 12-24.
- Stichting Benchmark GGZ (2017). Minimale Dataset: Factsheet Meetinstrumenten SCL-90 20170101. https://www.sbggz.nl/MDS.
- 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.

SUMMARY

ROM, benchmark and risk adjustment in a private practice

R.M. MOSTERMAN

BACKGROUND The benchmark methods of routine outcome monitoring (ROM)-data by the Dutch Mental Health Care Benchmark Foundation (SBG) have not been evaluated in independent research. No benchmark studies concerning private practices have been published. The Alliance of Quality in Mental Health Care (AkwaGGZ) has taken over the development of ROM; an excellent opportunity for improvement.

AIM To replicate (parts of) the SBG studies in a solo private practice.

METHOD Observational research on repeated measures with the SCL-90.

RESULTS During more than 12 years nearly 1,200 SCL-90 questionnaires have been completed by 644 patients. The ROM-group (ffl2 measures, n = 339) and non-ROM-group (n = 280) differed in patient and treatment characteristics, but no differences in pre-test were found. In the ROM-group 73% reliable improvement was found and 45% clinically recovered. For the purpose of benchmarking, the ROM-cohort was divided in three subgroups, representing three so-called providers. Benchmark with these providers revealed differences in progress and results. Post-test was predicted by pre-test, co-morbidity, low and medium education, which explained 33% of the variance. Adjustment for these variables increased the discrepancies between providers.

CONCLUSION Applying ROM in clinical practice is, at patient level, a useful tool to manage treatment process and determine its termination. Within-therapist-benchmark of aggregated measurements is helpful in detecting patterns and pitfalls. The SBG studies contain methodological imperfections. Risk variables cannot be generalized over samples and risk adjustment should be avoided.

TIJDSCHRIFT VOOR PSYCHIATRIE 62(2020)1, 27-36

KEYWORDS benchmark, private practice, risk adjustment, ROM