KORTE BIJDRAGE

Routine outcome monitoring en benchmarking: hoe kunnen we behandelresultaten op een zorgvuldige manier vergelijken?

M.J. NOOM, K. DE JONG, B. TIEMENS, F. KAMSTEEG, M.T. MARKUS, A.M. POT, G.M. SCHIPPERS, W. SWILDENS, S. SYTEMA, J. THEUNISSEN, H.P. VAN DER VLIST, J. VUYK, T. ZONDERVAN

ACHTERGROND *Het structureel meten van de resultaten van een behandeling in de geestelijke gezondheidszorg en het vergelijken daarvan tussen instellingen helpen om inzicht te krijgen in het effect van behandelingen in de reguliere praktijk.*
DOEL *Geven van een overzicht van de kwesties die van belang zijn bij het vergelijken van instellingen.*
METHODE *Analyseren van documentatie en beleidsinformatie over en praktijkervaring met routine outcome monitoring (ROM).*
RESULTATEN *We beschrijven knelpunten die kunnen ontstaan bij het vergelijken van instellingen en formuleren oplossingsrichtingen voor deze knelpunten. Daarbij staat centraal dat het werken met ROM een groeiproces is, waarbij men experimenteert met verschillende oplossingsrichtingen en op basis van ervaringen definitieve keuzes maakt.*
CONCLUSIE *Het is leerzaam om instellingen te vergelijken, zowel onderling als met 'best practices' (benchmarking). Instellingen verschillen echter in cliëntenpopulaties, meetprocedures en instrumentarium. Een zinvolle vergelijking is op termijn toch mogelijk.*
[TIJDSCHRIFT VOOR PSYCHIATRIE 54(2012)2, 141-145]

TREFWOORDEN benchmarking, geestelijke gezondheidszorg, routine outcome monitoring, vergelijkbaarheid

GGZ Nederland is in 2009 een project gestart om een landelijke vorm van routine outcome monitoring (ROM) in te voeren. Het doel van dit project is om te stimuleren dat ROM meer toegepast wordt in de ggz voor:
- behandelen en begeleiden;
- leren;
- verantwoorden;
- onderzoek.

GGZ Nederland heeft een werkgroep 'Vergelijkbaarheid' ingesteld om ideeën uit te werken over het zo zorgvuldig mogelijk vergelijken van behandelresultaten. De auteurs van dit artikel zijn lid van deze werkgroep. In dit artikel gaan wij specifiek in op de aandachtspunten bij het maken van vergelijkingen voor benchmarking.

Camp (1989) definieert benchmarking als *'systematisch onderzoek naar de prestaties en de onderliggende processen en methoden van één of meer leidende referentieorganisaties op een bepaald gebied, en de vergelijking van de eigen prestaties en werkmethoden met deze 'best practices', met het doel om de eigen prestaties te*

plaatsen en te verbeteren'. Het uitgangspunt is dat door vergelijken de mogelijkheid ontstaat van elkaar te leren. Bij de vergelijkbaarheid van gegevens moeten we echter aandacht schenken aan verschillen tussen doelgroepen, cliëntenpopulaties, ROM-procedures en ten slotte meetinstrumenten. Deze zullen wij achtereenvolgens bespreken.

VERGELIJKBAARHEID VAN DOELGROEPEN

Bij de prestatie-indicatoren (Zichtbare Zorg GGZ 2009) is gekozen voor een indeling van uitkomsten gebaseerd op diagnostische categorieën. De werkgroep heeft gekozen voor een indeling in doelgroepen. De reden hiervoor is dat instellingen meestal naar doelgroep georganiseerd zijn en dat meetinstrumenten voor ROM gekoppeld zijn aan de behandeldoelstellingen die bij een bepaalde doelgroep horen. De volgende doelgroepen zijn onderscheiden:

- volwassenen in kortdurende zorg;
- volwassenen met ernstige psychiatrische aandoeningen;
- ouderen;
- kinderen en jeugd;
- verslavingszorg;
- forensische zorg.

Een probleem met deze indeling is dat sommige stoornissen, zoals autismespectrumstoornissen, maar ook angst- en stemmingsstoornissen, in meerdere doelgroepen voorkomen. Toch lijkt deze indeling in doelgroepen bruikbaar om te komen tot vergelijkbare groepen. Een vervolgstap kan zijn om binnen een doelgroep onderscheid te maken naar diagnostische categorieën.

VERGELIJKBAARHEID VAN INSTELLINGEN

Binnen een doelgroep kunnen gegevens van instellingen met elkaar vergeleken worden. De vraag daarbij is hoe vergelijkbaar cliëntenpopulaties, ROM-procedures en meetinstrumenten zijn.

Cliëntenpopulaties

Instellingen verschillen in de demografische gegevens en klinische problematiek van hun clienten, de *casemix* (France e.a. 2001). Verschillen tussen instellingen in uitkomsten kunnen liggen aan verschillen tussen hun cliëntenpopulaties in plaats van verschillen in kwaliteit van zorg. Omdat we van veel doelgroepen niet weten welke factoren van invloed zijn, is casemixcorrectie voorlopig slechts beperkt mogelijk. Daarbij neemt men alleen factoren op die daadwerkelijk voorspellend zijn voor de uitkomsten (Zaslavsky 2001). Door de komende jaren naast uitkomstgegevens ook achtergrondvariabelen van cliënten en behandelgegevens te verzamelen, kunnen we meer inzicht krijgen in relevante factoren.

ROM-*procedures*

Instellingen verschillen ook in hun ROM-procedures. Zo is er variatie in het tijdstip van de metingen en de wijze van afname: wie de meting afneemt (een administratief medewerker, een behandelaar of een testverpleegkundige), hoe (op papier, via een interview of via een computer) en waar (thuis of bij de instelling). Bij langdurende zorg is er vaak geen sprake van een begin- of eindmeting, maar worden zorg en metingen gecontinueerd. Voor benchmarking zouden de procedures idealiter gelijk moeten zijn. Om behandelresultaten goed te kunnen vergelijken is het raadzaam dat beginmetingen zo vroeg mogelijk en eindmetingen zo laat mogelijk in de behandeling plaatsvinden. Voorts zou men op den duur ook follow-upmetingen kunnen verrichten, om na te gaan of de resultaten na een periode zonder zorg beklijven.

Meetinstrumenten

Er is veel discussie over meetinstrumenten, al lijkt er steeds meer consensus te ontstaan over welke het geschiktst zijn. In het landelijke ROM-

project is ervoor gekozen om uit te gaan van een selectie van gangbare instrumenten in de praktijk en dus verschillende instrumenten toe te staan. De commissie onderzoekt hoe gegevens onderling vergelijkbaar gemaakt kunnen worden. Het verdient de voorkeur om methodes te kiezen die ook toepasbaar zijn op individuele cliënten, zodat eenvoudig de verbinding gelegd kan worden tussen behandelresultaten op individueel niveau en op instellingsniveau.

De eenvoudigste methode daarvoor is *standaardisatie van scores*, waardoor elk instrument dezelfde schaal krijgt. Een voorbeeld hiervan zijn T-scores (De Beurs 2010). Een nadeel van standaardisatie is dat het geen criterium geeft voor de hoeveelheid verandering die noodzakelijk is om toeval uit te sluiten.

De *reliable change index* (RCI; Jacobson & Truax 1991) geeft wel een criterium voor het minimale verschil tussen twee metingen om van een betrouwbare verandering te kunnen spreken, maar zegt niets over de klinische betekenis van uitkomsten. Bij kortdurende zorg is de gewenste behandeluitkomst bijvoorbeeld herstel, terwijl dit bij langdurende zorg niet altijd nagestreefd wordt.

In de kortdurende zorg pas men vaak het principe van *klinische significantie* (Jacobson & Truax 1991) toe. Daarbij kijkt men hoe de eindmeting zich verhoudt tot de normscore voor 'normaal' functioneren. In de langdurende zorg zijn er methodes voorgesteld om te bepalen hoeveel personen over tijd naar een betere 'categorie' gaan, bijvoorbeeld van ernstig naar minder ernstig (o.a. Parabiaghi e.a. 2005). Ook kijkt men wel naar remissie.

Bij de doelgroep forensische psychiatrie is risicotaxatie van belang, maar er lijken nog geen normen te bestaan voor risicoreductie.

We kunnen concluderen dat er voor de meeste doelgroepen dus nog criteria voor betekenisvolle verandering ontwikkeld moeten worden. Hierin ligt een taak voor de doelgroepgerichte expertgroepen van GGZ Nederland.

RESULTATEN VAN PILOTSTUDIE

De werkgroep heeft een pilotstudie uitgevoerd bij 11 ggz-instellingen met 1500 cliënten uit vier verschillende doelgroepen (Tiemens e.a. 2010). We hebben in de doelgroep 'volwassenen in kortdurende behandeling' eerst vergelijkingen gemaakt met de RCI. Er waren geen significante verschillen tussen instellingen, maar wel tussen diagnostische groepen (zie tabel 1; χ^2 (*6*) = 13,4; p = 0,04).

Uit de pilotstudie blijkt hoe groot de invloed is van cliëntvariabelen: het is dus van belang daarvoor te corrigeren. Bovendien valt er veel te verbeteren in de aanlevering van de gegevens, zoals de hoofddiagnose en de berekening van som- of schaalscores van verschillende instrumenten.

TABEL 1 *Verschil in reliable change index* (RCI; *in %) bij verschillende stoornissen*

klinisch herstel en RCI	*Stemmings-stoornis (n = 135)*	*Angst-stoornis (n = 118)*	*Overige stoornissen (n = 187)*
Hersteld	30	21	35
Betrouwbaar verbeterd	19	17	11
Geen verbetering	42	54	50
Betrouwbaar verslechterd	8	8	4

BESCHOUWING

Het vergelijken van instellingen bij benchmarking is een delicaat proces, waarbij men zorgvuldig moet omgaan met verschillen tussen instellingen. Dat is nog meer van belang als benchmarking wordt ingezet voor externe verantwoording en financiële consequenties kan hebben. Daarbij is er een spanningsveld tussen algemene wensen die gelden voor ROM voor benchmarking en specifieke wensen voor een instelling om ROM 'op maat' te organiseren. Bovendien moet men een start maken met het registreren van type interventie, intensiteit en behandelduur. Deze informatie is cruciaal om uiteindelijk echt uitspraak te kunnen doen over wat werkt, voor wie en waarom. Ten slotte is het goed om te realiseren dat

een belangrijke voorwaarde voor een zorgvuldige toepassing van ROM is dat de kwaliteit van de verzamelde gegevens goed is. Een 'experimentele' periode is noodzakelijk, waarin men leert uit ervaringen met ROM in de praktijk en waarin men de besproken methoden verder ontwikkelt (Tiemens e.a. 2010). In deze periode kan men aan de hand van de database toetsen of de minimale dataset aanpassing behoeft, welke criteria voor verandering zinvol zijn en welke instrumenten het geschiktst zijn.

Naarmate meer gegevens verzameld zijn, kan men op empirische gronden keuzes maken op genoemde punten en zijn steeds nauwkeuriger vergelijkingen mogelijk tussen diagnosecategorieën, doelgroepen, procedures en meetinstrumenten. Dan kunnen uit de landelijke database referentiegroepen samengesteld worden die passen bij de casemix van een bepaalde instelling. Als een instelling zich bijvoorbeeld richt op de behandeling van kortdurende depressie en relatief veel allochtone cliënten heeft, relatief veel vrouwen en relatief veel ouderen, dan kan er een referentiegroep gecreëerd worden, waar deze instelling zich aan kan spiegelen. De keuze voor relevante variabelen voor het samenstellen van een referentiegroep moet gebaseerd zijn op resultaten van onderzoek. Een andere mogelijkheid is om op basis van de casemix de verwachte uitkomsten voor een instelling te bepalen en de instelling met haar eigen 'verwachte' uitkomst te benchmarken. Dan wordt benchmarking 'leren door vergelijken' en geen 'afrekenen voor financiering'.

CONCLUSIE

Wij concluderen dat er voldoende technieken beschikbaar zijn om ROM-gegevens met elkaar te vergelijken. Door zorgvuldig om te gaan met de mogelijkheden en de beperkingen, kan men inzicht verkrijgen in de mate waarin behandelresultaten kunnen overeenkomen of verschillen.

LITERATUUR

Beurs E de. De genormaliseerde T-score. Een euro voor testuitslagen. Maandblad Geestelijke volksgezondheid 2010; 65: 684-95.

Camp RC. Benchmarking: The search for industry best practices that lead to superior performance. Milwaukee: Quality press for the American society for quality control: 1989.

France FHR, Mertens I, Closon M, Hofdijk J. Case mix – Global view, local actions: Evolution in twenty countries. Amsterdam: IOS Press; 2001.

Jacobson N, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol 1991; 59: 12-9.

Parabiaghi A, Barbate AD, Avanzo B, Erlichter A, Lora A. Assessing reliable and clinically significant change on health of the nation outcome scales: Method for displaying longitudinal data. Aust N Z J Psychiatry 2005; 39: 719-25.

Tiemens B, Noorthoorn E, Janssen W, Kloos M. ROM GGZ, een pilot: ervaringen, mogelijkheden en aandachtspunten. Amersfoort: GGZ Nederland; 2010.

Zaslavsky AM. Statistical issues in reporting quality data: Small samples and casemix variation. Int J Qual Health Care 2001; 13: 481-8.

Zichtbare Zorg ggz. Basisset prestatie-indicatoren 2009-2010 geestelijke gezondheidszorg en verslavingszorg. Den Haag: Zichtbare zorg ggz; 2009.

AUTEURS

MARC NOOM universitair hoofddocent, Leiden Institute for the Advancement and Integration of Routine Outcome Monitoring (LIAIROM), LUMC, Leiden.

KIM DE JONG, onderzoeker, verbonden aan GGZ NHN en Erasmus Medisch Centrum, Rotterdam.

BEA TIEMENS, andragoog en epidemioloog, Pro Persona Centre for Education and Science, directeur Stichting Centrum voor Zorgmonitoring, en hoofd Onderzoek Indigo Service Organisatie.

FRANS KAMSTEEG, hoofd kenniscentrum GGNet, Warnsveld.

MONICA MARKUS, psycholoog-methodoloog en coördinator Routine Outcome Monitoring Consortium Kind- & Adolescentie Psychiatrie (ROMCKAP), Curium-LUMC, Oegstgeest.

ANNE MARGRIET POT, hoogleraar Ouderenpsychologie, Vrije Universiteit, Amsterdam en Trimbos-instituut, Utrecht.

GERARD SCHIPPERS, bijzonder hoogleraar Verslavingsgedrag en Zorgevaluatie, Amsterdam Institute for Addiction Research, AMC en Arkin, Amsterdam.
WILMA SWILDENS, senior onderzoeker en projectleider Utrechtse zorgmonitor, Altrecht, Utrecht.
SJOERD SYTEMA, senior onderzoeker bij het Rob Giel Onderzoekscentrum (RGOc), verbonden aan het Universitair Medisch Centrum Groningen (UMCG).
JAN THEUNISSEN, senior onderzoeker bij GGZ inGeest, partner van VUmc, Amsterdam.
PAUL VAN DER VLIST, business consultant, EQ Groep, Woerden.
JUDITH VUYK, psychiater-directeur, Willem Arntsz Divisie, Altrecht, Utrecht.
TINEKE ZONDERVAN, hoofd afdeling Nieuwe Kennis, Delta Psychiatrisch Centrum, Rotterdam.
Correspondentieadres: dr. Marc Noom, LUMC, Afdeling Psychiatrie, Albinusdreef 2, 2333 ZA Leiden.
E-mail: m.j.noom@lumc.nl

Geen strijdige belangen meegedeeld.

Het artikel werd voor publicatie geaccepteerd op 28-11-2011.

SUMMARY

Routine outcome monitoring and benchmarking: how can treatment results be compared in a responsible manner? M.J. Noom, K. de Jong, B. Tiemens, F. Kamsteeg, M.T. Markus, A.M. Pot, G.M. Schippers, W. Swildens, S. Sytema, J. Theunissen, H.P. van der Vlist, J. Vuyk, T. Zondervan –

BACKGROUND *The structural measurement of the results of treatment under the Dutch mental health services and a comparison of these results between mental health centres help to provide insight into the effectiveness of treatment in general practice.*

AIM *To provide an overview of the issues that require attention when the results of mental health centres are being compared.*

METHOD *Documentation, policy information and practical experience with routine outcome monitoring were analysed.*

RESULTS *We describe the problems that can arise when results obtained by mental health centres are compared and we suggest some solutions for these problems. Important factors that have emerged from our study are as follows: working with routine outcome monitoring is a process of natural growth and involves experiences with several solutions and the making of definitive choices on the basis of experience.*

CONCLUSION *It is instructive to compare mental health centres with each other and with regards to so-called 'best practices' (benchmarking). However, mental health centres draw on a differing wide mix of patients and use different measurement procedures and instruments. In this article we express the view that in the near future it should be possible to draw meaningful comparisons.*

[TIJDSCHRIFT VOOR PSYCHIATRIE 54(2012)2, 141-145]

KEY WORDS benchmarking, comparison, mental health care, routine outcome monitoring