Edwin de Beurs

Hoofd wetenschappelijk onderzoek SBG

Hoogleraar ROM en Benchmarken Universiteit Leiden

Naar aanleiding van het rapport van de Algemene Rekenkamer en STOPROM

Inleiding

De discussie over het nut van ROM en Benchmarken in de GGZ is sinds het rapport van de Algemene Rekenkamer (AR) van januari 2017 (Algemene Rekenkamer, 2017) opgelaaid. Tegenstanders zien bureaucratische last, subjectieve metingen, onbetrouwbare gegevens en een te grote bemoeienis van de zorgverzekeraars. Voorstanders wijzen op de voordelen van ROM voor het behandelproces, vinden dat ze wat kunnen leren van uitkomstgegevens op groepsniveau en achten het redelijk dat de sector verantwoording aflegt voor het geld dat jaarlijks in de GGZ omgaat.

De conclusies in het rapport van de AR gingen overigens niet over ROM, maar over de huidige vorm van prestatiebekostiging in de GGZ en over de mogelijkheden van *uitkomst*bekostiging in de GGZ. De AR concludeert dat er anno 2017 belangrijke stappen zijn gezet met het vergroten van de transparantie in de GGZ, maar dat er van bekostiging op basis van kwaliteit van zorg nog niet veel terecht is gekomen. Onderhandelingen tussen zorgaanbieders en zorgverzekeraars zijn vooral gericht op beheersing van het budget en de kwaliteit van de geleverde zorg speelt hierbij nog geen rol. Verder adviseert de AR om voor de invoering van “uitkomstbekostiging” ruim de tijd te nemen. Voor een samenhangend en effectief bekostigingsmodel moet nog ervaring worden opgedaan met zorgstandaarden en de kwaliteit van de ROM-gegevens acht de AR nu nog onvoldoende; de gegevens zijn nog niet volledig genoeg en nog niet voldoende vergelijkbaar. Tegenstanders van ROM zien hier aanleiding in om te pleiten voor het stoppen met kwaliteitsmetingen op basis van ROM. De nuance in de discussie is soms behoorlijk zoek. Daarom een uitgebreide reactie op de commentaren die gegeven zijn.

ROM ≠ Benchmarken ≠ Verantwoorden ≠ Afrekenen

In de discussie lijkt ROM te staan voor alles wat er mis is in de GGZ: toenemende administratieve last, dwang om nutteloze vragenlijsten af te nemen, bemoeienis van de verzekeraars met de inhoud van de zorg. In de argumentatie wordt vaak ROM, Benchmarken, verantwoorden en afrekenen op één hoop gegooid. Zo kan het dat de oproep tot het tekenen van de petitie STOPROM vergezeld gaat van de aanbeveling niet met ROM te stoppen, maar wel te stoppen met misbruik door de verzekeraars van de ROM-gegevens voor “benchmarken”, waar dan “afgerekend worden op je resultaten” mee bedoeld lijkt te worden. Deze begripsverwarring helpt de discussie niet verder. ROM is niet hetzelfde als benchmarken en dat is niet hetzelfde als verantwoorden en dat is weer wat anders dan afrekenen op uitkomsten door de zorgverzekeraar. Ik zet eerst uiteen wat met die verschillende begrippen bedoeld wordt en ga dan in op een aantal stellingen die in de discussie naar voren worden gebracht.

Wat is ROM?

ROM staat voor Routine Outcome Monitoring en houdt in dat gedurende de behandeling regelmatig gemeten wordt hoe het met de patiënt gaat. Zo houdt de behandelaar een vinger aan de pols bij de patiënt, bespreekt de meetresultaten met de patiënt en ze kunnen zo samen de koers van de behandeling bepalen. In de geneeskunde is dit al langer gebruikelijk (denk aan laboratoriumuitslagen, beeldvormende technieken zoals Röntgen of (f)MRI scans, of het bepalen van uitzaaiingen in de

oncologie). Ook in de GGZ wordt al lang gemeten - hier vooral klachten, symptomen en functioneren - maar pas sinds 2010 is ROM op grote schaal ingevoerd. Dat is geen unieke Nederlandse ontwikkeling (Roe, Drake, & Slade, 2015). In het buitenland zien we dezelfde ontwikkelingen, bijvoorbeeld in Engeland met het Improved Access to Psychotherapy project (IAPT; Clark et al., 2009), in Australië en Nieuw-Zeeland waar men de HoNOS inzet bij ernstige psychiatrische aandoeningen (Burgess, Pirkis, & Coombs, 2015) en in de VS met verscheidene projecten, zoals uitkomstmeting van depressiebehandeling van alle zorgaanbieders in Minnesota (zie www.mnhealthscores.org). Een aantal van deze initiatieven werden in 2015 door internationale sprekers toegelicht op de conferentie van SBG.

Wat is Benchmarken?

Benchmarken is een strategie voor kwaliteitsverbetering. Door op groepsniveau uitkomstgegevens te bestuderen om ervan te leren kun je de kwaliteit van zorg verbeteren (Ettorchi-Tardy, Levif, & Michel, 2012). Door je eigen uitkomsten te vergelijken met anderen kan je praktijkvariatie en best practices op het spoor komen, waarmee je je eigen praktijkvoering kunt verbeteren (Barendregt, 2015). Daar worden goede ervaring mee opgedaan (McKay, Coombs, & Duerden, 2014; Roe et al., 2016). In de industrie en in de gezondheidszorg is dit een uitermate succesvolle strategie gebleken om tot kwaliteitsverbetering te komen (Camp, 1989; Poerstamper, van Mourik - van Herk, & Veltman, 2007).

Wat is verantwoorden?

Verantwoorden is laten zien wat de GGZ oplevert. Als sector wil je kunnen laten zien dat de ruim vier miljard die jaarlijks in de curatieve GGZ omgaat, goed wordt besteed. Verantwoorden kan op macro-, meso-, en microniveau: de gehele GGZ, een instelling of afdeling of de caseload van een individuele behandelaar. Verantwoorden is belangrijk aangezien in de samenleving het beeld van de GGZ als black box regelmatig terugkeert, omdat onvoldoende helder is wat er in de spreekkamer van de behandelaar gebeurt en wat behandeling oplevert. Verantwoorden kan op allerlei manieren (bijvoorbeeld visitaties in het kader van certificering, laten zien dat je volgens de richtlijnen werkt, dat je voldoet aan nascholingseisen, wetenschappelijk effectiviteitsonderzoek in de klinische praktijk van alledag, et cetera), maar de meest directe wijze van verantwoorden is om te laten zien in hoeverre patiënten baat hebben bij behandeling.

Wat is afrekenen?

Afrekenen (of uitkomstenbekostiging) is instellingen belonen voor het leveren van goede kwaliteit en ‘straffen’ voor het leveren van slechte kwaliteit. In de onderzoeksliteratuur wordt dit “pay-for-performance” (P4P) genoemd. Het Nederlandse systeem van gereguleerde marktwerking is hier met de invoering van de nieuwe zorgverzekeringswet van 2006 voor ingericht (Eijkenaar & Schut, 2015). Sinds 2006 is een systeem van prestatiebekostiging ingevoerd. Dit houdt in dat betaald wordt per verrichting, in de GGZ per Diagnose-Behandel-Combinatie (DBC). Een mogelijke volgende stap is de invoering van uitkomstenbekostiging of P4P. Er is in Nederland nog weinig ervaring opgedaan met P4P in de gezondheidszorg en de internationale ervaringen zijn ook niet onverdeeld gunstig. Extra financiering beloven voor betere kwaliteit heeft nog niet overtuigend tot verbetering van die kwaliteit geleid (Eijkenaar, Emmert, Scheppach, & Schöffski, 2013; Petersen, Woodard, Urech, Daw, & Sookanan, 2006). Eijkenaar en Schut (2015) schetsen de voorwaardelijke condities voor P4P en concluderen voorzichtig dat adequate P4P een begaanbare weg is, maar dat invoering per 2020 - het oorspronkelijke voornemen - niet mogelijk is. In de VS experimenteert men al langer met P4P-initiatieven (Burwell, 2015) en is het een belangrijk onderdeel van de Affordable Care Act (beter bekend als Obamacare). Hier worden wel positieve ervaringen gerapporteerd voor de algemene gezondheidszorg en ook een review over P4P in de GGZ is voorzichtig positief: “results suggest that P4P can lead to improved quality and efficiency” (Stewart, Lareef, Hadley, & Mandell, 2016). Met een meta-analyse is aangetoond dat er veel heterogeniteit is in de evaluatie van P4P; het resultaat is afhankelijk van de stoornis, de uitkomstmaat (proces of outcome) en de methodologische kwaliteit van de studie (Ogundeji, Bland, & Sheldon, 2016). Over de zegeningen van P4P en van marktwerking in

de zorg is het laatste woord nog niet gezegd en lopen de opinies uiteen. Hoe het ook zij, wanneer er financiële consequenties aan de kwaliteit van zorg verbonden gaan worden is goede informatie over die kwaliteit nodig.

Ik wil hieronder een aantal stellingen die in de huidige discussie over ROM langskomen kort aanstippen en nuanceren. Voor een uitgebreidere bespreking van ROM en Benchmarken verwijs ik naar mijn oratie van november 2015 (de Beurs, 2015).

"Ik kan zelf wel zien hoe het gaat met mijn patiënt"

Helaas is dit maar ten dele het geval. Bij het inschatten van het klinische beeld van een patiënt zijn er al snel grenzen aan de hoeveelheid informatie die we tegelijk kunnen overzien (ongeveer 7 aspecten; Miller, 1956). Al in de jaren '50 van de vorige eeuw heeft Meehl aangetoond dat een meting met een gestandaardiseerd instrument superieur is aan het oordeel van een clinicus (Grove & Meehl, 1996; Meehl, 1954). Een meetinstrument is dus een welkome ondersteuning bij deze complexe taak. Clinici blijken hun eigen vaardigheid om tot een betrouwbaar klinisch oordeel te komen te overschatten (Mohr, 1995). Verder blijkt dat behandelaars vaak een te gunstig beeld hebben van hoe het met hun patiënten gaat en niet goed zien aankomen dat hun patiënt verslechtert (Hannan et al., 2005). Hannan en collega's doen verslag van onderzoek onder 550 patiënten; 40 patiënten verslechterden gedurende de behandeling, maar slechts bij één patiënt werd dit voorspeld door de behandelaar, terwijl de verslechtering met de OQ-45 bij 36 van de 40 patiënten kon worden voorspeld. Een goede test levert dus op zijn minst een belangrijke bijdrage ter ondersteuning van het klinisch oordeel.

"ROM is ongeschikt om kwaliteit van zorg te meten"

Kwaliteit van zorg meten is ingewikkeld. Jaarlijks is er onder medici veel discussie als Elsevier's Weekblad de lijst publiceert van ziekenhuizen, gerangschikt naar kwaliteit van zorg. Donabedian heeft het kwaliteitsdenken in de zorg een flinke impuls gegeven (Donabedian, 1980). Hij onderscheidt drie aspecten waaraan de kwaliteit van zorg afgemeten kan worden: wat je doet (of het proces, zoals de inhoud van de behandeling, de duur, de vorm), waarmee je het doet (of zogenaamde structuur factoren, zoals het jaarlijks budget van de zorgaanbieder, het opleidingsniveau van de behandelaar, het gebouw, de "cultuur" van de zorgaanbieder) en wat de zorg heeft opgeleverd (of de uitkomstfactoren, zoals pre-post vermindering van klachten of symptomen, verbetering van functioneren, welbevinden, patiëntervaring). Voor al deze factoren zijn er meetinstrumenten en alle aspecten zijn relevant. Kwaliteit verbeteren gaat niet zonder te meten (Berwick, James, & Coye, 2003; Lazar, Fleischut, & Regan, 2013). Gedacht vanuit de patiënt gaat het er vooral om wat de zorg oplevert in termen van gezondheidswinst. Daarmee is de belangrijkste indicator van de kwaliteit van zorg de uitkomst van de behandeling volgens de patiënt (Porter & Teisberg, 2006). Porter noemt dit "patient-based value" van gezondheidszorg. Dat er meer is aan kwaliteit dan alleen de uitkomst van de behandeling staat buiten kijf; dat de kwaliteit alleen zou zijn af te meten aan structuur- en procesfactoren is niet waar.

"De patiënt heeft hier niks aan"

ROM-metingen voor en na de behandeling, alleen om te kunnen aanleveren bij SBG en daarmee te voldoen aan de eisen van de zorgverzekeraar, is niet direct nuttig voor de patiënt. Goede ROM houdt in dat je veel frequenter meet om te zien of de patiënt goed reageert op de behandeling. Dat kan door tussenmetingen uit te voeren of door continue te meten zoals met het Feedback Informed Treatment (FIT; Miller, Duncan, Sorrell, & Brown, 2005) of zoals men meet in het Improved Access to Psychotherapy project (IAPT; Clark et al., 2009). De feedback die ROM oplevert kan op die manier ook een wezenlijk onderdeel uitmaken van gezamenlijke besluitvorming of Shared Decision Making (Metz et al., 2015) en we kunnen zo de zorg beter op de behoeften van patiënt afstemmen of meer "patient-centered" maken. Dit geeft veel meer regie aan patiënten waardoor ze meer bij de behandeling betrokken raken. Onderzoek van Lambert en collega's heeft aangetoond dat de ROM variant waarbij zowel de patiënt als de behandelaar feedback krijgt tot betere resultaten leidt dan

wanneer alleen de behandelaars feedback krijgt of wanneer feedback geheel uitblijft (Lambert, Harmon, Slade, Whipple, & Hawkins, 2005). Het is dus essentieel om de resultaten van iedere ROM meting te bespreken met de patiënt en het behandelbeleid erop af te stemmen. Alleen dan komt de waarde van ROM ten volle tot zijn recht. Een nog op te lossen kwestie is of generieke meetinstrumenten voldoende gedetailleerde informatie bieden om de behandelaar en patiënt goed te informeren of dat we daar juist stoornisspecifieke instrumenten voor moeten gebruiken (Blankers, Barendregt, & Dekker, 2016). De bestaande ROM-infrastructuur biedt in ieder geval ruimte om ook stoornisspecifieke of klachtspecifieke instrumenten aan een ROM batterij toe te voegen. ROM is opgenomen in de zorgstandaarden van het Netwerk Kwaliteitsontwikkeling (zie www.kwaliteitsontwikkelingggz.nl/) en ook in de meerjaren ontwikkelagenda voor de GGZ “Agenda voor gepast gebruik en transparantie”, die door de koepelorganisatie voor patiënten LPGGZ, GGZ-Nederland en alle betrokken beroepsorganisaties is opgesteld, wordt het belang van ROM (voor de zorgverlening aan de patiënt én voor transparantie over de zorg) onderschreven (VWS, 2015).

“Uitkomsten zijn subjectief”

In het rapport van de Algemene Rekenkamer komen we dit als kritiekpunt op ROM tegen: het gaat om “subjectieve” metingen, door de patiënt ingevulde beoordelingen van zichzelf en geen “harde” objectieve feiten zoals bloedwaarden, post-operatieve infecties of sterfte. De auteurs van het rapport lijken te menen dat ROM-metingen niets zeggen over de objectieve baat die men heeft van een behandeling in de GGZ. In ROM maken we gebruik van gestandaardiseerde en gevalideerde meetinstrumenten. Dit levert geobjectiveerde metingen op van psychische klachten, die in essentie subjectieve fenomenen zijn. Angst, depressie, desoriëntatie, (somatisch onbegrepen) lichamelijke klachten zijn niet te zien in een (f)MRI-scanner of op een röntgenfoto, maar kunnen we alleen meten door de patiënt erover te bevragen. Een ander voorbeeld is pijn: ook niet objectief aantoonbaar maar zelf-rapporteerde pijn is wel de belangrijkste uitkomstmaat om behandeling voor pijn te evalueren (Dworkin et al., 2005). Als het waar zou zijn dat “subjectieve” vragenlijsten geen goed beeld geven van het effect van behandelingen in de GGZ, dan verkruimelt de evidence base onder de psychiatrie en de klinische psychologie, want gecontroleerde onderzoeken maken bijna altijd gebruik van de meetmethodiek die ook bij ROM wordt toegepast. Overigens zien we in de somatische zorg ook steeds meer interesse voor ROM. ROM heet daar PROM voor Patient Reported Outcome Measurement (Black, 2013; Dawson, Doll, Fitzpatrick, Jenkinson, & Carr, 2010). Met ROM loopt de GGZ als sector dus voorop op andere sectoren in de gezondheidszorg en vanuit daar heeft men volop belangstelling voor de ROM-infrastructuur die in de GGZ tot stand is gekomen.

Uniformiteit

Dat de één de temperatuur uitdrukt in graden Celsius en de ander in Fahrenheit of dat de één afstanden meet in kilometers en de ander in mijlen komt een heldere uitwisseling van informatie niet ten goede en kan leiden tot een Babylonische spraakverwarring. Een enkele taal voor kwaliteit is zeer gewenst en standaardisering en uniformiteit is noodzakelijk. Natuurlijk is iedere patiënt en iedere behandeling uniek, maar het daarbij laten brengt de wetenschap en de klinische praktijk niet verder. We moeten ook kijken naar gedeelde kenmerken van patiënten en gemeenschappelijke elementen van behandeling. Classificatiesystemen voor psychische aandoeningen zoals de DSM helpen daarbij, evenals richtlijnen voor behandeling zoals opgenomen in de zorgstandaarden van het Netwerk Kwaliteitsontwikkeling GGZ (NKO). Dit kan leiden tot een goede taxonomie (een systematisch classificatiesysteem) voor therapeutische interventies in de GGZ en dat is wenselijk (Bradley, Curry, & Devers, 2007). Een taxonomie helpt om complexe fenomenen zoals het therapeutisch handelen in de GGZ te verhelderen en scherper te definieerden (Sofaer, 1999). Voor uitkomstmeting hebben we de T-score in de GGZ geïntroduceerd (de Beurs, 2010). Gebruik van een uniforme meeteenheid voor de ernst van de psychopathologie vergemakkelijkt de communicatie tussen professionals. Uiteraard kan de T-score ondersteund worden door andere (uitkomst)maten die een indicatie geven van de kwaliteit van zorg. Zoals ik hierboven al aangaf is kwaliteit meer dan alleen een Delta T.

De huidige verworvenheden vieren en de pijnpunten oplossen

Sinds 2010 is er in de GGZ heel wat bereikt. Er is een infrastructuur tot stand gebracht om te meten, er zijn flinke stappen gezet om te komen tot uniformiteit in meetinstrumenten en meetmomenten, Pijnpunten zijn er ook nog: ROM implementeren louter en alleen om respons-percentages te halen, leidt tot een bureaucratische variant van ROM waar de behandelaar en de patiënt weinig aan hebben en waar men terecht tegen in het verweer komt. ROM dient in de eerste plaats de behandeling te ondersteunen. Bruikbare ROM hoeft echter helemaal niet op gespannen voet te staan met ROM voor kwaliteitsmanagement. Individuele ROM-gegevens kunnen geaggregeerd worden en bieden zicht op wat behandeling in de GGZ op groepsniveau oplevert. Zo komen “best practices” boven drijven en leren we wat het best werkt bij wie onder welke omstandigheden (Paul, 1967).

Van STOPROM naar verbeter ROM

Vaak blijft ROM beperkt tot een voor- en nameting die alleen worden uitgevoerd om gegevens bij SBG aan te kunnen leveren en zo de responsafspraken na te komen die met de zorgverzekeraar zijn gemaakt. Dat is nooit de bedoeling van ROM geweest en ik ben het volledig eens met de kritiek die de initiatiefnemers van STOPROM hebben op deze toepassingsvorm van ROM (zie ook de Beurs, 2015). ROM wordt pas waardevol voor het primaire proces wanneer je vaker meet en de meetresultaten inzet bij de behandeling. Laten we in plaats van te stoppen met uitkomstmetingen op basis van ROM (STOPROM) de nationale informatie infrastructuur die de afgelopen jaren in de GGZ tot stand is gebracht nog beter gaan benutten door de huidige ROM te verbeteren. Beter meten betekent frequenter en breder meten en - waar nodig - met de inzet van stoornisspecifieke meetinstrumenten. Zo krijgen we betere informatie om het klinische proces te ondersteunen en ook betere informatie om de kwaliteit van de zorg te monitoren en te verbeteren. Met betere zorg zijn onze patiënten beter af en dat is de ultieme motivator waarvoor iedere professional in de GGZ zou moeten gaan.

Van SBG 1.0 naar SBG 2.0

De kwaliteit van GGZ meten is ingewikkeld en houdt niet op bij het bepalen van de uitkomst van de zorg; aan kwaliteit zitten vele kanten en de kwaliteit van een ziekenhuis of instelling terugbrengen tot een positie op een ranglijst schiet hoe dan ook te kort en kan ongewenste bijeffecten hebben. Veel gehoorde suggesties om het meten van uitkomsten in de GGZ verder te verbeteren zijn: breder meten dan symptomen en functioneren door ook te kijken naar persoonlijke doelstellingen van de patiënt; klinisch betekenisvolle uitkomsten presenteren (% verbeterde of herstelde patiënten). SBG werkt voortdurend aan verbetering van de benchmarkmethodiek en laat zich daarbij adviseren door een wetenschappelijke raad en diverse expertraden met leden uit het GGZ-veld. Zo hebben we statistische correcties voor casemix verschillen tussen instellingen ontwikkeld (Warmerdam, Barendregt, & de Beurs, 2017) en criteria voor betrouwbare verandering en klinisch significant herstel ingebouwd in de benchmark (de Beurs et al., 2016) om de zeggingskracht van de uitkomstgegevens te vergroten. Een andere veel gehoorde wens is niet alleen te kijken naar de uitkomst onmiddellijk na beëindiging van behandeling, maar ook naar de uitkomst op lange termijn te kijken. Landelijk follow-up gegevens verzamelen is echter geen sinecure.

SBG is “werk-in-uitvoering”

Zoals de wetenschappelijke raad van SBG eerder stelde “SBG is werk-in uitvoering” (Blijd-Hoogewys et al., 2012) en verbeteren van de benchmark methodiek is altijd mogelijk. Verbeteren is iets alledaags en op alles van toepassing. In de gezondheidszorg is een Hiv-infectie van een doodvonnis veranderd in een chronische aandoening, waar je oud mee kan worden; in de oncologie zijn op de patiënt toegesneden behandelingen effectief, waar voorheen de genezingskans nihil was. Ook in de GGZ behandelen we nu anders dan 20 jaar geleden en gaan we dezelfde weg op (staging en profiling, zie: Beekman, van Os, van Marle, & van Harten, 2012). Kortom, we verbeteren voortdurend alles. ROM en Benchmarken zijn beiden technieken om verbetering van zorg te bewerkstelligen (de Beurs, 2015). En we kunnen er meteen mee zien of we dat doel daadwerkelijk bereiken. Wat voor de gezondheidszorg

en de GGZ geldt, geldt natuurlijk ook voor SBG. Over vijf of tien jaar hebben we beter zicht op de kwaliteit van de zorg dan nu. Dat lukt echter alleen met inzet van alle betrokkenen: patiënten, ROM-medewerkers, professionals, managers, bestuurders en wetenschappers.

Referenties

Algemene Rekenkamer. (2017). *Bekostiging van de curatieve geestelijke gezondheidszorg*. Retrieved from Den Haag:

Barendregt, M. (2015). Benchmarken en andere functies van ROM: back to basics. *Tijdschrift voor Psychiatrie, 57*(7), 517-525. Retrieved from www.tijdschriftvoorpsychiatrie.nl/assets/articles/57-2015-7-artikel-barendregt.pdf

Beekman, A. T., van Os, J., van Marle, H. J., & van Harten, P. N. (2012). Stagering en profilering van psychiatrische stoornissen [Staging and profiling of psychiatric disorders]. *Tijdschrift voor Psychiatrie, 54*(11), 915-920. Retrieved from http:/europepmc.org/abstract/MED/23138617

Berwick, D. M., James, B., & Coye, M. J. (2003). Connections between Quality Measurement and Improvement. *Medical Care, 41*(1), I30-I38. Retrieved from http://www.jstor.org/stable/3767726

Black, N. (2013). Patient reported outcome measures could help transform healthcare. *British Medical Journal, 346*, f167. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23358487

Blankers, M., Barendregt, M., & Dekker, J. J. M. (2016). Meetvariatie als bron van bias bij het benchmarken met verschillende ROM-instrumenten. *Tijdschrift voor Psychiatrie, 58*(1), 55-66. Retrieved from www.tijdschriftvoorpsychiatrie.nl/en/issues/497/articles/10752

Blijd-Hoogewys, E., van Dijck, R., Emmelkamp, P., Mulder, N., Oude Voshaar, R. C., Schippers, G., . . . Vermeiren, R. (2012). Benchmarken is 'werk-in-uitvoering'. *Tijdschrift voor Psychiatrie, 54*(12), 1031-1038. Retrieved from www.ncbi.nlm.nih.gov/pubmed/23250645

Bradley, E. H., Curry, L. A., & Devers, K. J. (2007). Qualitative Data Analysis for Health Services Research: Developing Taxonomy, Themes, and Theory. *Health Services Research, 42*(4), 1758-1772. doi:10.1111/j.1475-6773.2006.00684.x

Burgess, P., Pirkis, J., & Coombs, T. (2015). Routine outcome measurement in Australia. *International Review of Psychiatry, 27*(4), 264-275. doi:10.3109/09540261.2014.977234

Burwell, S. M. (2015). Setting value-based payment goals--HHS efforts to improve US health care. *New England Journal of Medicine, 372*(10), 897-899. doi:10.1056/NEJMp1500445

Camp, R. C. (1989). *Benchmarking: the search for industry best practices that lead to superior performance*. Wisconsin: ASQC Quality Resources.

Clark, D. M., Layard, R., Smithies, R., Richards, D. A., Suckling, R., & Wright, B. (2009). Improving access to psychological therapy: Initial evaluation of two UK demonstration sites. *Behaviour Research and Therapy, 47*(11), 910-920. doi:10.1016/j.brat.2009.07.010

Dawson, J., Doll, H., Fitzpatrick, R., Jenkinson, C., & Carr, A. J. (2010). The routine use of patient reported outcome measures in healthcare settings. *British Medical Journal, 340*, 186. doi:10.1136/bmj.c186

de Beurs, E. (2010). De genormaliseerde T-score, een 'euro' voor testuitslagen [The normalised T-score: A euro for test results]. *Maandblad Geestelijke Volksgezondheid, 65*, 684-695. Retrieved from www.sbggz.nl

de Beurs, E. (2015). *ROM en Benchmarken, over meten, weten en wat dan? (oratie)*. Leiden: Leiden University.

de Beurs, E., Barendregt, M., de Heer, A., van Duijn, E., Goeree, B., Kloos, M., . . . Merks, A. (2016). Comparing methods to denote treatment outcome in clinical research and benchmarking mental health care. *Clinical Psychology & Psychotherapy, 23*, 308-318. doi:10.1002/cpp.1954

Donabedian, A. (1980). *The definition of quality and approaches to its assessment*. Ann Arbor, Mich.: Health Administration Press.

Dworkin, R. H., Turk, D. C., Farrar, J. T., Haythornthwaite, J. A., Jensen, M. P., Katz, N. P., . . . Witter, J. (2005). Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. *Pain, 113*(1-2), 9-19. doi:10.1016/j.pain.2004.09.012

Eijkenaar, F., Emmert, M., Scheppach, M., & Schöffski, O. (2013). Effects of pay for performance in health care: A systematic review of systematic reviews. *Health Policy, 110*(2), 115-130. doi:10.1016/j.healthpol.2013.01.008

Eijkenaar, F., & Schut, E. (2015). *Uitkomstbekostiging in de zorg: een (on)begaanbare weg?* Retrieved from Rotterdam: http://repub.eur.nl/pub/78057/

Ettorchi-Tardy, A., Levif, M., & Michel, P. (2012). Benchmarking: A method for Continuous Quality Improvement in Health. *Healthcare Policy, 7*(4), e101-e119. doi:PMC3359088

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law, 2*(2), 293-323. doi:10.1037/1076-8971.2.2.293

Hannan, C., Lambert, M. J., Harmon, C., Nielsen, S. L., Smart, D. W., Shimokawa, K., & Sutton, S. W. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology in Medical Settings, 61*(2), 155-163. doi:10.1002/jclp.20108 [doi]

Lambert, M. J., Harmon, C., Slade, K., Whipple, J. L., & Hawkins, E. J. (2005). Providing feedback to psychotherapists on their patients' progress: clinical results and practice suggestions. *Journal of Clinical Psychology, 61*(2), 165-174. doi:10.1002/jclp.20113

Lazar, E. J., Fleischut, P., & Regan, B. K. (2013). Quality measurement in healthcare. *Annual Review of Medicine, 64*, 485-496. doi:10.1146/annurev-med-061511-135544

McKay, R., Coombs, T., & Duerden, D. (2014). The art and science of using routine outcome measurement in mental health benchmarking. *Australian and Asian Psychiatry, 22*(1), 13-18. doi:10.1177/1039856213511673

Meehl, P. E. (1954). *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Minneapolis University of Minnesota.

Metz, M. J., Franx, G. C., Veerbeek, M. A., de Beurs, E., van der Feltz-Cornelis, C. M., & Beekman, A. T. F. (2015). Shared Decision Making in mental health care using Routine Outcome Monitoring as a source of information: a cluster randomised controlled trial. *BMC Psychiatry, 15*(1), 1-10. doi:10.1186/s12888-015-0696-2
Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81-97. doi:10.1037/h0043158
Miller, S. D., Duncan, B. L., Sorrell, R., & Brown, G. S. (2005). The partners for change outcome management system. *Journal of Clinical Psychology, 61*(2), 199-208. doi:10.1002/jclp.20111
Mohr, D. C. (1995). Negative outcome in psychotherapy: A critical review. *Clinical Psychology: Science and Practice, 2*(1), 1-27. doi:10.1111/j.1468-2850.1995.tb00022.x
Ogundeji, Y. K., Bland, J. M., & Sheldon, T. A. (2016). The effectiveness of payment for performance in health care: A meta-analysis and exploration of variation in outcomes. *Health Policy, 120*(10), 1141-1150. doi:10.1016/j.healthpol.2016.09.002
Paul, G. L. (1967). Strategy of outcome research in psychotherapy. *Journal of Consulting Psychology, 31*(2), 109-118. doi:10.1037/h0024436
Petersen, L. A., Woodard, L. D., Urech, T., Daw, C., & Sookanan, S. (2006). Does pay-for-performance improve the quality of health care? *Annals of Internal Medicine, 145*(4), 265-272. doi:10.7326/0003-4819-145-4-200608150-00006
Poerstamper, R.-J., van Mourik - van Herk, A., & Veltman, A. (2007). *Benchmarking in Dutch health care: Towards an excellent organisation*. Amsterdam: PricewaterhouseCoopers.
Porter, M. E., & Teisberg, E. O. (2006). *Redefining Health care: creating value-based competition on results*. Cambridge: Harvard Business Press.
Roe, D., Drake, R. E., & Slade, M. (2015). Routine outcome monitoring: An international endeavour. *International Review of Psychiatry, 27*(4), 257-260. doi:10.3109/09540261.2015.1070552
Roe, D., Lapid, L., Baloush-Kleinman, V., Garber-Epstein, P., Gornemann, M. I., & Gelkopf, M. (2016). Using Routine Outcome Measures to Provide Feedback at the Service Agency Level. *Community Mental Health Journal, 52*(8), 1022-1032. doi:10.1007/s10597-016-0039-x
Sofaer, S. (1999). Qualitative methods: what are they and why use them? *Health Services Research, 34*(5 Pt 2), 1101-1118. doi:PMCID: PMC1089055
Stewart, R. E., Lareef, I., Hadley, T. R., & Mandell, D. S. (2016). Can we pay for performance in behavioral health care? *Psychiatric Services, 68*(2), 109-111. doi:10.1176/appi.ps.201600475
VWS, M. v. (2015). *Agenda ggz voor gepast gebruik en transparantie* (25 424). Den Haag: Sdu Retrieved from https://www.rijksoverheid.nl/documenten/rapporten/2015/11/26/agenda-ggz-voor-gepast-gebruik-en-transparantie.
Warmerdam, E. H., Barendregt, M., & de Beurs, E. (2017). Risk adjustment of self-reported clinical outcomes in Dutch mental health care. *Journal of Public Health*. doi:10.1007/s10389-017-0785-4