Rating scales in treatment efficacy studies

Hafkenscheid, Antonius Joseph Petrus Maria

SAMENVATTING

HET GEïNDIVIDUALISEERDE EN NORMATIEVE GEBRUIK VAN PSYCHIATRISCHE BEOORDELINGSSCHALEN IN THERAPEUTISCH EFFECTONDERZOEK

Bezwaren tegen het klassieke effectonderzoek
Resultaten, verkregen in psycho- en farmacotherapeutisch effectonderzoek worden van oudsher geanalyseerd en geïnterpreteerd met behulp van de klassieke (toetsende) statistiek. Uitspraken omtrent de (vergelijkende) werkzaamheid van therapieën op basis van het statistische significantiecriterium zijn om verschillende redenen beperkt of zelfs misleidend.

Een van de belangrijkste bezwaren tegen het gebruik van statistische significantietoetsen in therapeutisch effectonderzoek betreft de wijze, waarop met interindividuele verschillen wordt omgegaan. Scores op effectmaten worden per behandelconditie gemiddeld, aangezien -analoog aan de principes van het gecontroleerd experimenteel onderzoek- interindividuele verschillen binnen de proefgroep(en) als "foutenvariantie" worden opgevat. Profiteren patiënten (proefpersonen) in de proefgroep(en) ongeveer evenveel (of weinig) van een bepaalde therapie (met andere woorden: is er weinig verschil in effectiviteit tussen patiënten binnen een proefgroep), dan kunnen *gemiddelde* verschillen in werkzaamheid ten opzichte van concurrerende therapieën (of ten opzichte van een placebobehandeling) niettemin statistische significantie bereiken.

Het omgekeerde is eveneens mogelijk. Is een bepaalde experimentele behandeling bijvoorbeeld uiterst effectief bij enkele patiënten (proefpersonen), terwijl dezelfde behandeling ineffectief of zelfs schadelijk blijkt te zijn voor andere patiënten, dan verschilt bij nameting het gemiddelde van de experimentele groep mogelijk te weinig ten opzichte van het gemiddelde van de (vrijwel onveranderde patiënten in de) controlegroep om de klassieke nulhypothese te kunnen verwerpen. De werkzaamheid van de therapie bij een (klein) *deel* van de patiënten kan echter wel degelijk klinisch betekenisvol zijn.

In de praktijk van de geestelijke gezondheidszorg zijn aanzienlijke interindividuele verschillen in de mate waarin een bepaalde therapie effect sorteert eerder regel dan uitzondering, zelfs in homogene patiëntengroepen met een identieke psychiatrische diagnose.

Zulke interindividuele verschillen worden uitsluitend zichtbaar door behandeleffecten *per afzonderlijke patiënt* te analyseren. Uit dergelijke intraindividuele analyses kunnen vervolgens (per proefgroep en per effectmaat) *percentages* effectief behandelde -of anders geformuleerd: klinisch significant verbeterde- patiënten worden gegenereerd. Zo'n benadering heeft in het algemeen meer informatiewaarde dan het louter rapporteren van *gemiddelde* (verschillen in) effectiviteit.

Het classificeren van *afzonderlijke* patiënten als effectief behandeld -oftewel: klinisch significant verbeterd- veronderstelt dat

1) er algemeen aanvaarde, eenduidige en repliceerbare normen beschikbaar zijn om te kunnen bepalen bij welke scores of scoreveranderingen er sprake is van een effectieve behandeling;
2) een score*verandering* minimaal de onbetrouwbaarheidsmarges van de effectmaat overstijgt.

De indexen voor klinische significantie en betrouwbare verandering
Ongeveer tien jaar geleden stelden Jacobson, Follette & Revenstorf (1984), in een inmiddels klassiek geworden artikel, drie indexen voor die klinisch significante behandeleffecten, elk op hun eigen wijze, normatief definiëren. In datzelfde artikel introduceerden zij een index voor het bepalen van betrouwbare verandering; dat wil zeggen: verandering, die de onbetrouwbaarheidsmarges van de effectmaat overstijgt.

De *eerste* index voor klinische significantie is gebaseerd op de vergelijking van individuele scores bij nameting met het gemiddelde van een psychiatrische (dysfunctionerende) normgroep. Een klinisch significant behandeleffect is gedefinieerd als een score bij nameting, die tenminste twee standaarddeviaties buiten het bereik van het gemiddelde voor een psychiatrische (dysfunctionerende) normgroep valt, in de richting van "gezond" functioneren.

De *tweede* index stelt dat uitsluitend scores die na behandeling minder dan twee standaarddeviaties afwijken van het gemiddelde voor een "gezonde" normgroep als klinisch significant gelden.

Het *derde* voorstel definieert een score bij nameting als klinisch significant, als de afwijking van het gemiddelde voor een "gezonde"normgroep kleiner is dan de afwijking ten opzichte van het gemiddelde voor een psychiatrisch (dysfunctionerende) normgroep.

De indexen voor klinische significantie en voor betrouwbare verandering zijn in principe algemeen toepasbaar, ongeacht de aard en inhoud van de onderzochte therapievorm(en). Tot dusver heeft de belangstelling voor de voorstellen van Jacobson et al. (1984), die in twee latere publicaties (Jacobson & Revenstorf, 1988; Jacobson & Truax, 1991) werden herhaald, zich echter grotendeels beperkt tot het gedragstherapeutisch effectonderzoek. In gedragstherapeutische kringen worden de indexen beschouwd als de tot dusver meest uitgewerkte en minst arbitraire definities van behandeleffectiviteit.

Vragen ten aanzien van de indexen
Niettemin laten de voorgestelde indexen nog veel vragen open. Deze vragen zijn deels empirisch, deels theoretisch van aard. Wat betreft de drie definities van klinische significantie is een *empirische* vraag, in hoeverre de drie indexen resulteren in verschillende schattingen van behandeleffectiviteit. *Theoretisch* is bijvoorbeeld de vraag aan de orde, of en zo ja welke uitsluitingscriteria (het al dan niet in psychiatrische of psychotherapeutische behandeling zijn, het al dan niet behalen van een "pathologische" score op een effectmaat etcetera) gehanteerd moeten worden bij de samenstelling van een "gezonde" normgroep.

Wat betreft de index voor betrouwbare verandering: inmiddels zijn zes *theoretische* varianten of correcties bekend op de formule, die door Jacobson et al. (1984) is voorgesteld. Een *empirische* vraag is in welke mate schattingen van percentages betrouwbaar veranderde patiënten op basis van de zeven beschikbare indexen voor betrouwbare verandering verschillen.

Een betrouwbaarheidsschatting van de gebruikte effectmaat is in alle formules voor betrouwbare verandering onontbeerlijk. De vraag in hoeverre percentages betrouwbaar veranderde patiënten uiteenlopen bij verschillende betrouwbaarheidsschattingen (in een en dezelfde formule voor betrouwbare verandering) is *empirisch* van aard.
Of de betrouwbaarheidsschatting gebaseerd dient te zijn op de interne consistentie, test-hertestbetrouwbaarheid of (in geval van gedragsobservatieschalen) interbeoordelaarsbetrouwbaarheid is een onderwerp van *theoretische* discussie.

Tot slot is de (empirische) vraag van belang, in hoeverre schattingen van klinisch significante behandeleffecten en van betrouwbare verandering afhankelijk zijn van de gebruikte effectmaten, vooral als die effectmaten ongeveer dezelfde meetpretentie hebben.

Deze (en andere) vragen zijn het onderwerp van onderhavige dissertatie.
De drie indexen voor klinische significantie en de zeven indexen voor betrouwbare verandering worden voor drie effectmaten empirisch vergeleken aan de hand van de opname- en ontslagscores van 121 kortdurend opgenomen Nederlandse psychiatrische patiënten.

De drie effectmaten vertegenwoordigen elk een eigen perspectief op de manifeste psychopathologie (het toestandsbeeld) van de patiënt:

1) de gestandaardiseerde versie van de "Brief Psychiatric Rating Scale" (BPRS-18 Overall & Gorham, 1962; Lukoff, Liberman & Nuechterlein, 1986) is een beoordelingsschaal, die door *klinici* wordt afgenomen;
2) De "Nurses' Observation Scale for Inpatient Evaluation" (NOSIE-30; Honigfeld Gillis & Klett, 1966) is een *verpleegkundige* gedragsobservatieschaal en
3) de "Symptom Checklist-90" (SCL-90; Derogatis, Lipman & Covi, 1973), is een *zelfbeoordelings*schaal voor psychopathologie.

Elk in hun soort behoren deze beoordelingsschalen tot de meest bekende en gebruikte effectmaten in de internationale vakliteratuur. Hun psychometrische eigenschappen (betrouwbaarheidsschattingen, validiteitsaspecten) zijn evenwel, zeker voor de Nederlandstalige versies, nog onvoldoende bekend.

De psychometrische voorstudies
De dissertatie bestaat uit twee delen. Deel 1 is gewijd aan eigen psychometrisch onderzoek naar de drie beoordelingsschalen in verschillende Nederlandse steekproeven. In hoofdstuk 2 wordt de gestandaardiseerde BPRS-18 psychometrisch geëvalueerd. Hoofdstuk 3 is een replicatie van het onderzoek naar de interbeoordelaarsbetrouwbaarheid van de BPRS-18. Daarnaast wordt in dat hoofdstuk de temporele betrouwbaarheid (stabiliteit) van beoordelingen met de BPRS bepaald. Hoofdstuk 4 is een kritische bespreking van methode

om de interbeoordelaarsbetrouwbaarheid van de BPRS en aanverwante beoordelingsschalen te maximaliseren. Met name worden kanttekeningen geplaatst bij de validiteit van beoordelingen door meerdere klinici in een en hetzelfde gesprek met de patiënt, alsmede bij het gebruik van op videoband vastgelegde "status praesens" vraaggesprekken. In hoofdstuk 5 wordt de -vrijwel volledig impliciet gebleven- verdeeldheid die er binnen de onderzoeksgemeenschap heerst ten aanzien van de "juiste" methode om de interbeoordelaarsbetrouwbaarheid van "interview-based" beoordelingsschalen zoals de BPRS te bepalen nogeens kort geaccentueerd. Hoofdstuk 6 is een verslag van een uitgebreid psychometrisch onderzoek van de NOSIE-30, waarbij ondermeer interbeoordelaarsbetrouwbaarheid, test-hertestbetrouwbaarheid, interne consistentie en dimensionele structuur werden bepaald. In hoofdstuk 7 worden de interne consistentie, temporele betrouwbaarheid en dimensionele structuur van de SCL-90, waarvan de psychometrische eigenschappen vrijwel uitsluitend bekend waren bij *ambulante* psychiatrische patiënten, onderzocht in een grote, heterogene, recent *opgenomen* psychiatrische patiëntengroep.

In het algemeen komen de verschillende betrouwbaarheidsaspeckten van de drie effectmaten (interne consistentie en test-hertestbetrouwbaarheid voor alle drie effectmaten, interbeoordelaarsbetrouwbaarheid voor BPRS-18 en NOSIE-30) op *schaal*niveau als redelijk tot bevredigend naar voren. Alhoewel de oorspronkelijke dimensionele structuur voor twee van de drie beoordelingsschalen (BPRS-18 en NOSIE-30) goed repliceerbaar blijkt, zijn de interbeoordelaarsbetrouwbaarheden op *sub*schaalniveau doorgaans onvoldoende. De dimensionele structuur van de SCL-90 blijkt slechts matig repliceerbaar.

Uit deze resultaten wordt geconcludeerd dat een multidimensioneel gebruik van de drie effectmaten niet gerechtvaardigd is, maar dat hun totaalscores wel bruikbaar zijn in therapeutisch effectonderzoek.

Het hoofdonderzoek
Deel 2 behelst de eigenlijke vragen van dit dissertatie-onderzoek: hoofdstuk 8 de empirische, hoofdstuk 9 de theoretische. In hoofdstuk 8 wordt onderzocht, in hoeverre schattingen van behandeleffectiviteit variëren met (a) het gebruik van verschillende indexen voor klinische significantie, (b) op basis van de gebruikte index voor betrouwbare verandering en (c) afhankelijk van de gebruikte effectmaat.

Gemiddeld (groepsgewijs) is de daling in manifeste psychopathologie tussen opname en ontslag voor elk van de drie effectmaten statistisch significant, waarbij het statistisch effect "middelgroot" (NOSIE-30) of "groot" (BPRS-18, SCL-90) is. De schattingen van klinisch significante behandeleffectiviteit blijken sterk afhankelijk van de gebruikte index voor klinische significantie. Op basis van de tweede index (een behandeleffect is klinisch significant als de score bij nameting minder dan twee standaarddeviaties afwijkt van het gemiddelde van een "gezonde" normgroep) worden hoge schattingen van klinisch significante behandeleffectiviteit verkregen. Deze schattingen hebben evenwel nauwelijks betekenis, omdat volgens deze definitie vrijwel alle patiënten eveneens een *voor*metingsscore behalen, die binnen het "normale" bereik valt. Dit is overigens mede toe te schrijven aan de (pragmatisch en theoretisch verdedigbare) keuze van de "gezonde" normgroepen: deze waren samengesteld uit *ambulante* psychiatrische patiënten (en niet uit

een afspiegeling van de gewone bevolking). De scoreverdelingskenmerken van d "gezonde" en/of psychiatrische (dysfunctionerende) normgroepen maken het gebruik of d interpretatie van de eerste en derde index voor klinische significantie (vrijwel) onmogelijk De schattingen van betrouwbare verandering blijken (sterk) afhankelijk te zijn van (a) d gebruikte index voor betrouwbare verandering, (b) de gebruikte betrouwbaarheidsschattin (interne consistentie, test-hertestbetrouwbaarheid, interbeoordelaarsbetrouwbaarheid) en (c de gebruikte effectmaat.

De drie indexen voor klinische significantie zijn feitelijk alleen gebaseerd op scores b *na*meting. Om te kunnen spreken van een klinisch significante *verbetering* dient er tever (minimaal) sprake te zijn van een *betrouwbare* verbetering.
De schattingen van behandeleffectiviteit op basis van scores en scoreveranderingen die aa *beide* voorwaarden (een klinisch significante score bij nameting èn een scoreverbeterin die de onbetrouwbaarheidsmarges van de effectmaat overstijgt) voldoen zijn tamelijk te zeer bescheiden: veelal minder dan 50% van de patiënten in de steekproef.

Een theoretische beschouwing van de indexen:
In hoofdstuk 9 worden enkele theoretische vooronderstellingen, die aan de indexen voo klinische significantie en betrouwbare verandering ten grondslag liggen kritisch besproke Achtereenvolgens worden kanttekeningen geplaatst bij: (1) de aanname, dat scoreve delingen voor "gezonde" en/of psychiatrische normgroepen normaalverdelingen ben deren; (2) de aanname, dat patiënten en "gezonden" "van nature" tot *afzonderlijk* normgroepen behoren; (3) de in- en uitsluitingscriteria, die (impliciet of expliciet) bij d samenstelling van de normgroep(en) worden gehanteerd;
(4) het gebruik van tweedelingen om patiënten te classificeren als (al dan niet) klinisc significant verbeterd en/of betrouwbaar veranderd; (5) de waardeoordelen, die in d definities van klinische significantie besloten liggen en (6) het gebruik van verschilscore tussen voor- en nameting om klinisch significante en betrouwbare verandering te meten.

Samenvattend moet worden concludeerd dat de indexen voor klinisch significantie e betrouwbare verandering klinisch en theoretisch een zekere aantrekkingskracht uitoefene maar dat hun (empirische) toepassing en (theoretische) interpretatie allerminst probleer loos zijn.

LITERATUUR

Derogatis L.R., Lipman R.S., Covi L. (1973). SCL-90, an outpatient psychiatric ratin scale-preliminary report. *Psychopharmacology Bulletin, 9*, 13-28.

Honigfeld G., Gillis R.D., Klett C.J. (1966). NOSIE-30: a treatment sensitive wa behaviour scale. *Psychological Reports, 19*, 180-182.