Assessment of change in clinical evaluation

Middel, Lambertus Johannes

8

Samenvatting

Dit proefschrift is het resultaat van een aantal artikelen waarin, met behulp van verschillende ragenlijstmethoden, verandering in gezondheidstoestand is gemeten. Het doel is na te gaan wat de betekenis van zo'n verandering in gezondheidstoestand is. Hoofdstuk 1 is een inleiding op de wijze waarop onderzoekers vat proberen te krijgen op de vraag: in hoeverre zijn statistisch significante verschillen ook relevante verschillen, d.w.z. ook substantiële verschillen. Immers, in tegenstelling tot kleine steekproeven, hebben triviale verschillen in grote steekproeven een grotere kans om statistisch significant zijn. Een veel gebruikte methode om de relevantie van verschillen te kwantificeren en te interpreteren, wordt in de literatuur aangeduid met het begrip 'responsiveness'. Het concept responsiveness wordt niet eenduidig gehanteerd. In studies waarin nieuwe instrumenten worden gevalideerd wordt vaak met responsiveness 'de gevoeligheid van een meetinstrument om verandering te meten' bedoeld als een psychometrische eigenschap van het instrument. Om aan te tonen dat het ene meetinstrument meer gevoelig is verandering te meten dan een soortgenoot (bijvoorbeeld een ziekte-specifiek versus generiek) is het begrip 'relatieve responsiveness' geïntroduceerd. In studies waarin (medische) interventies op werkzaamheid worden geëvalueerd, wordt het begrip responsiveness ook gebruikt maar dan als indicator voor ook de grootte van het aan de behandeling gerelateerde effect. Bij dit type onderzoek is dus de relevantie van het gevonden verschil tussen voor- en nameting bij klinische interventies bepaald door de hoogte van de effect size. Een groot effect (effect size > .80) wordt dan als een klinisch relevant of klinisch belangrijk verschil wordt geïnterpreteerd door onderzoekers. De vraag is of dit terecht is. Om een verschil als klinisch belangrijk te kunnen interpreteren is een extern criterium nodig voor die belangrijkheid of relevantie. Om verschillen of veranderingen in gezondheidstoestand die gemeten zijn met verschilscores verkregen door longitudinale meting met vragenlijsten te kunnen interpreteren als klinisch relevant of belangrijk, is in dit proefschrift het oordeel van de patiënt als extern criterium gebruikt. Een dergelijk criterium wordt doorgaans gevormd door een score op retrospectieve vraag aan de patiënt in welke mate hij of zij vindt onveranderd, verbeterd of verslechterd te zijn na de behandeling.

In dit proefschrift is, met alle items van een vragenlijst verandering in de gezondheidstoestand longitudinaal gemeten en is, na de interventie met dezelfde items in een retrospectieve vorm, de gepercipieerde verandering gemeten met transitie-items. Er is nog veel onbekend over de relatie tussen de longitudinale verschilscore die de onderzoeker gebruikt om te kunnen oordelen over de richting en grootte van de verandering, en de retrospectieve transitie scores op schaal-items

waarmee het oordeel van de patiënt over de richting en omvang van de verandering wordt gemeten. Tevens is er zeer weinig bekend over de psychometrische eigenschappen van meetinstrumenten die in retrospectieve vorm de gepercipieerde verandering meten. Het tweede doel van dit proefschrift is om na te gaan in welke mate beide oordelen met elkaar overeenstemmen, welke over- en onderschattingen van effectgrootte er gemaakt kunnen worden bij afhankelijke steekproeven en welke psychometrische eigenschappen het meetinstrument in retrospectieve vorm heeft.

In hoofdstuk 2 is worden de veranderingen bestudeerd in een gerandomiseerde studie bij een groep patiënten waarbij oraal toegediende medicatie in maximale doses geen resultaat meer gaf in de verwachtte vermindering van ernstige vormen van spasticiteit. Door de medicatie intrathecaal toe te dienen, door middel van een subcutaan geïmplanteerde programmeerbare pomp, werd een substantiële vermindering van spasticiteit en verbetering in gezondheidstoestand verwacht. Om de verandering tussen baseline en post-test als ook de verschillen in verandering tussen de groep intrathecaal toegediende placebo en de groep met werkzame stof (Baclofen) te kunnen beoordelen, zijn non-parametrische toetsen en effect sizes toegepast. In deze studie zijn effect sizes alleen gepresenteerd indien de verschillen niet aan toevalsfluctuaties toe te schrijven zijn. De toedieningswijze van medicatie voor ernstige spasticiteit resulteerde na een jaar in statistisch significante en relevante verschillen in zowel klinische parameters als in scores op de lichamelijke dimensies van gezondheidstoestand gemeten met de Sickness Impact Profile en de Hopkins Symptoms Checklist. Deze studie illustreert het klassieke gebruik van statistische toetsen en effect sizes om daarmee uitspraken te doen over de relevantie van het gevonden effect.

In hoofdstuk 3 worden de psychometrische eigenschappen geëvalueerd van het meetinstrument dat in deze studie gekozen is om verandering in gezondheidstoestand te meten, de Nederlandse versie van de Minnesota Living with Heart Failure Questionnaire (MLHF-Q). De steekproef voor dit onderzoek betrof een groep patiënten die behandeld werd voor boezemfibrileren. De interne consistentie werd geschat met behulp van Cronbach's alpha; de construct validiteit werd beoordeeld met behulp van multitrait-multimethod analyse met gebruikmaking van schalen uit de Rand-36, uit de Hospital Anxiety and Depression Scale (HADS) en uit de Multidimensional Fatigue Inventory (MFI-20). Known-Group validiteit werd geëvalueerd door de verschillen te analyseren tussen de groep met de minst ernstige vorm van angina pectoris met de groep met de meer ernstige vorm aan de hand van het oordeel van een cardioloog (NYHA classificatie). De test-hertest betrouwbaarheid is geschat met een sub-groep in de steekproef waarvan de

gezondheidstoestand gedurende drie maanden na de baseline meting geen verandering had ondergaan. Om na te gaan of de dimensies van de factoranalyse uit deze steekproef, uit een Nederlandse populatie patiënten met een vorm van hartfalen, overeenkomen met die uit de oorspronkelijke factor analyse in de Amerikaanse studie, is gebruik gemaakt van Perfect Congruence Analysis (PCA). De MLHF-Q schalen hadden een bevredigende Cronbach's alpha en discrimineerden goed tussen de groepen met ernstige en minder ernstige angina pectoris. Het resultaat van de multitrait-multimethod analyse laat een goede overeenstemming zien tussen de 'fysiek functioneren' schaal van de MLHF-Q en de fysieke dimensies van de andere (concurrente) instrumenten. Voor de schaal 'emotioneel functioneren' is deze overeenstemming zwakker. De factor ladingen in de oorspronkelijke schaal constructie met een Amerikaanse steekproef kwamen goed overeen met de ladingen in de Nederlandse steekproef. De MLHF-Q bleek gevoelig om verandering te meten in een groep patiënten die een behandeling ondergaat die primair gericht is op verbetering van sinusritme. De test-hertest van de MLHF-Q resulteerde in bevredigende correlaties groter dan .60. De MLHF-Q is dus een betrouwbaar en valide instrument om veranderingen in gezondheidstoestand te meten bij patiënten met hartfalen.

In hoofdstuk 4 wordt de toepassing van de grenzen van Cohen om een effect size te interpreteren in termen van 'triviaal', 'klein', 'medium' of 'groot' nader geanalyseerd. Deze grenzen zijn door Cohen ontwikkeld op basis van een index gebaseerd op het verschil in gemiddelden tussen twee onafhankelijke steekproeven gedeeld door de gepoolde standaarddeviatie (SDp) Dit in eenheden van de gepoolde standaarddeviatie uitgedrukte verschil d (difference) is door Cohen geannoteerd als d'. Omdat Cohen in zijn standaardwerk over poweranalyse en effect size de tabellen waarmee de omvang van de steekproef te bepalen zijn heeft gebaseerd op twee steekproeven, corrigeert hij d' voor de situatie waarin er sprake is van bijvoorbeeld een herhaalde meting binnen 1 steekproef. Hiertoe corrigeert hij d' met $\sqrt{1-r}$ waarbij r de correlatie is tussen de meting op T1 en T2. Deze effect size is gelijk aan het gemiddelde verschil tussen T1 en T2 gedeeld door de standaarddeviatie van dat verschil. Deze effect size wordt veel gebruikt en staat bekend als de Standardised Response Mean (SRM). Daarnaast is er een aantal effect sizes ontwikkeld waar het gemiddelde verschil in eenheden van andere standaarddeviaties wordt uitgedrukt die van deze SDp en SRM afwijken. In dit hoofdstuk wordt aangetoond dat, als men de grenzen van Cohen, behorend bij de d', toepast op de SRM, er een risico is van over- en onderschatting van de geschatte effectgrootte in ongeveer één op de vijf in de gebruikte steekproef van effect sizes. Aangezien de SRM alleen te herleiden is tot d' als de correlatiecoëfficient tussen T1 en T2 berekend kon worden op grond van de

in de betreffende publicatie gerepresenteerd gegevens (148 van de 411 SRM's) heeft de analyse zich tot deze moeten beperken. Effect sizes indices geschat door het gemiddelde verschil te standaardiseren met de standaarddeviatie (SD) van o.a. de baseline scores (in subgroepen) of de SD van de veranderingsscores (in subgroepen) zijn wiskundig niet te herleiden tot de effect size d'. Derhalve is het kritiekloos toepassen van de grenzen van Cohen niet vrij van het risico van over- en onderschatting van de grootte van het effect.

In hoofdstuk 5 is de overeenkomst tussen het oordeel van de onderzoeker over de mate van verbetering (met behulp van effect size indices) en het oordeel van de patiënt over de mate van verbetering beschreven. De patiënten in deze steekproef (N=217) ondergingen een behandeling waarvan bekend is dat deze de gezondheidstoestand verbeterd. Twintig patiënten gaven na de behandeling aan verslechterd te zijn. De analyses zijn door dit kleine aantal beperkt tot patiënten die van mening waren te zijn verbeterd of onveranderd te zijn gebleven en na de interventie die zij ondergingen: Percutaneous Transluminal Coronary Angioplasty (PTCA), Coronary Artery Bypass Grafting (CABG) of farmacotherapie). In deze studie is de Nederlandse versie van de Minnesota Living with Heart Failure Questionnaire (MLHF-Q), aangevuld met enkele MOS-20 items, gebruikt om verandering in gezondheidstoestand te meten. Aangezien van elk item in de vragenlijst zowel een verschilscore als een bij dit item behorend retrospectief oordeel gemeten is, is in eerste instantie een vergelijking tussen onderzoekersoordeel (verschilscore) en patiëntoordeel (retrospectief oordeel) op item niveau uitgevoerd voor 23 items. Geen rekening houdend met de dimensies waartoe items behoren zijn 4798 response-combinaties berekend van een totaal van 4991 (217 x 23) als gevolg van 4% missing data.

Voor de bepaling van de concordantie tussen de gestandaardiseerde verschillen (effect grootte) van schalen of domeinen, is als extern criterium voor de relevantie van de verbetering in de schaalscores gebruik gemaakt van zogenoemde 'global questions' naar gepercipieerde verandering in deze domeinen. Een voorbeel van zo'n global question is 'Sinds de bypass operatie is mijn beperking in het trappenlopen: 1) sterk verbeterd, 2) nogal verbeterd, 3) weinig verbeterd, 4) onveranderd, 5) een beetje slechter, 6) nogal slechter en 7) sterk verslechterd'.
In dit hoofdstuk is geprobeerd een eenvoudig criterium te ontwikkelen om te bepalen wanneer het oordeel van de onderzoeker over de grootte van de verandering (effect size) in overeenstemming is met het externe criterium, of daarvan afwijkt in termen van een over- of onderschatting. De vuistregel van Cohen (< .20 triviaal; ≥ .20 < .50 klein; ≥ .50 < .80 medium; en ≥ .80 groot effect) waarmee effect grootte

doorgaans door onderzoekers worden geïnterpreteerd, lijken synchroon te lopen met het de oordelen van de patiënt in termen van gepercipieerde verbetering (respectievelijk: 'geen verandering'; 'een beetje verbeterd'; 'nogal verbeterd' en 'veel verbeterd'.

In hoofdstuk 6 is de relatie onderzocht tussen: 1) de gemeten verandering in de schalen 'lichamelijk' en 'emotioneel functioneren' met longitudinale meting en 2) de gepercipieerde verandering in deze domeinen. In deze studie is eveneens de Nederlandse versie van de Minnesota Living with Heart Failure Questionnaire (MLHF-Q), aangevuld met enkele MOS-20 items, gebruikt om verandering in gezondheidstoestand te meten. De gepercipieerde verandering in lichamelijk en emotioneel functioneren zijn gemeten met gemodificeerde schaal-items die opgenomen zijn in de vragenlijst die na de behandeling (Percutaneous Transluminal Coronary Angioplasty (PTCA), Coronary Artery Bypass Grafting (CABG) of farmacotherapie) is ingevuld. Deze modificatie bestaat uit de retrospectieve vorm van het betreffende item (transitie item genoemd) waarmee gevraagd werd naar de richting en de mate van verandering na de behandeling. Analoog aan de interne consistentie van de schaal van veranderingscores werd de interne consistentie van de transitie schalen geschat. De interne consistentie uitgedrukt in Cronbach's alpha van de met herhaalde meting verkregen veranderscores en die van de corresponderende transitieschaal 'fysiek functioneren' waren bevredigend. Factoranalyses van de baseline items, de verschilscore van deze items en de hiermee corresponderende transitie items resulteerden in dezelfde factorstructuur. Om onbetrouwbare 'eyeball' interpretatie te voorkomen zijn 95% betrouwbaarheidsintervallen rond de ladingen van de items binnen elke dataset berekend. De resultaten laten geen verschillen zien ondanks de verwachtte systematisch lagere ladingen van de item verschilscores vergeleken met die van de baseline items en transitie items. De canonische correlatie tussen de lineaire combinatie van de verschilscores per item en de lineaire combinatie van de transitie-items resulteerde voor beide schalen in 40% (verandering in fysiek functioneren) respectievelijk 23% (verandering in emotioneel functioneren) verklaarde variantie. De conclusie is dat beide meetmethoden overlap vertonen maar ook verschillende veranderingen registreren. In situaties waarin geen baselinemeting mogelijk is, zou met de transitiemethode volstaan kunnen worden.

In hoofdstuk 7 worden de belangrijkste resultaten van de voorgaande hoofdstukken samengevat en, op basis van de in een later stadium verkregen inzichten, kritisch geëvalueerd. De belangrijkst conclusies zijn dat het kritiekloos toepassen van de vuistregel van Cohen voor effect size op de Standardised Response Mean tot over en onderschattingen leidt. In hoofdstuk 4 is daarvoor een eenvoudige correctie

methode ontwikkeld. In een secundaire analyse van de gegevens uit de hoofdstukken 2 en 3 werden, in afwijking van het resultaat uit hoofdstuk 4, bijna twee maal zoveel over- of onderschattingen gevonden ten opzichte van Cohen's grenzen behorend bij effect size d'. De veel kleinere steekproef kan hier een rol hebben gespeeld. Een andere belangrijke conclusie is dat grootte van verschilscores verkregen uit herhaalde meting, in grote mate concordant zijn met de door de patiënt gepercipieerde grootte van de verandering. De in hoofdstuk 5 gevonden overeenstemming tussen het onderzoekers-oordeel over de effect grootte en de door de patiënt gepercipieerde verandering werd door een andere studie [1] bevestigd. De effect size in domeinen van de gepercipieerde verandering en die van de longitudinaal gemeten verandering blijken niet van elkaar te verschillen, zijn even groot voor groepen waar een gotere verandering verwacht wordt en waar deze niet verwacht wordt. Door de items uit een vragenlijst na de operatie af te nemen in de vorm van transitie-items en deze op schaalniveau te vergelijken met de veranderingscores, is aangetoond dat schalen die gepercipieerde veranderingen meten vergelijkbare effect sizes, factorladingen en interne consistentie hebben. De invloed van mogelijke confounders in de meting van gepercipieerde verandering in domeinen van gezondheid, is object van analyse en een publicatie na dit proefschrift. De beïnvloeding van het retrospectieve oordeel over het effect van behandeling door 'recall bias' (men weet niet meer in welke mate men voor de behandeling beperkt was in bijvoorbeeld trappenlopen) en de 'present state bias' (de mate van (niet) beperkt zijn op het moment van herhaalde meting na de behandeling) is onderwerp van lopend onderzoek.

---

1. Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. Journal of Clinical Oncology 1998;16(1):139-44.