Aan je eigen haren omhoog?

Over betrouwbaarheid en validiteit van instrumenten voor het meten van 'kwaliteit van leven'

JOOST ZAAT
FRANÇOIS SCHELLEVIS

Zaat J, Schellevis F. Aan je eigen haren omhoog? Over betrouwbaarheid en validiteit van instrumenten voor het meten van 'kwaliteit van leven'. Huisarts Wet 1995; 38(3): 105-9.

Samenvatting In dit artikel worden de keuzemomenten bij het kiezen van een geschikt meetinstrument voor het meten van effecten van medische behandelingen uitgelegd aan de hand van een fictief onderzoek. De keuze van het al dan niet meten van subjectieve aspecten hangt af van de klinische (en maatschappelijke) relevantie. Het doel bij het meten – discriminatie tussen groepen, prognostisch of evaluatief – heeft consequenties voor de aard van de vragen in het meetinstrument. Een meetinstrument moet betrouwbaar zijn: de vragen moeten met elkaar samenhangen (interne consistentie), bij onveranderde klinische toestand moet de uitkomst van de meting gelijk zijn (test-hertest-correlatie), terwijl er bij veranderingen juist een veranderde uitkomst verwacht wordt (responsiveness). Een instrument hoort ook te meten wat het moet meten (validiteit). Of de inhoud van de vragen overeenkomt met het te meten begrip valt onder inhoudsvaliditeit; de vergelijking van een nieuw meetinstrument met een gouden standaard (criteriumvaliditeit) is bij kwaliteit van leven instrumenten veelal niet mogelijk. Wel kunnen verschillende instrumenten met elkaar vergeleken worden of kan een instrument in verschillende stadia van een aandoening worden afgenomen (contructvaliditeit). Het valideren van een meetinstrument is meer een proces dat een statisch geheel.

Vakgroep Huisarts- en Verpleeghuisgeneeskunde / Instituut voor Extramuraal Geneeskundig Onderzoek Vrije Universiteit Amsterdam, Van de Boechorststraat 7, 1081 BT Amsterdam.
Dr. J.O.M. Zaat, Dr. F.G. Schellevis, beiden huisarts.
Correspondentie: Dr. J.O.M. Zaat.

Inleiding

The health status measurement literature is a jungle. Many get lost; some, presumably consumed by large carnivores, are never heard from again.[1]

Het aantal publikaties over het meten van 'kwaliteit van leven' in het kader van huisartsgeneeskundig onderzoek is groot. Een search in Medline op trefwoorden als 'functional status', 'health status' of 'quality of life' in combinatie met 'general practice/practitioner' of 'family practice' leverde alleen in 1993 al 46 publikaties op. Uit deze literatuur komt een zeer heterogeen beeld van kwaliteit van leven naar voren. Het is dan ook de vraag of het begrip kwaliteit van leven wel werkbaar is. *De Neeling* stelt voor om het voortaan over welzijnseffecten van medische behandelingen te hebben,[2] maar ook dat begrip dekt niet alles.

Aan deze effecten kun je een aantal niveaus te onderscheiden:
- de subjectieve klachten van de patiënt (jeuk, moeheid, pijn);
- het functioneren van de patiënt (functionele status) met lichamelijke, psychische, en sociale beperkingen als aandachtspunten (mobiliteit, stress, sociale rollen);
- de kwaliteit van leven in engere zin: het beeld dat een patiënt heeft van zijn eigen, geheel individuele kwaliteit van leven, in vergelijking met een door de patiënt zelf gekozen referentiekader.

Per definitie is er voor de kwaliteit van leven geen echte gouden standaard. De veel gebruikte meetinstrumenten richten zich op de 'functionele toestand' (bijvoorbeeld de 'Sickness Impact Profile') of het algemene gevoel van gezondheid (bijvoorbeeld de 'General Health Perception Questionnaire'). De literatuur over de verschillende instrumenten is uitgebreid.[3] Meetinstrumenten voor 'kwaliteit van leven' zijn voor een belangrijk deel gebaseerd op een psychometrisch begrippenkader. Voor de niet ingewijde lezer levert dat al snel een grote hoeveelheid onbegrijpelijke termen op.

We vroegen we ons af welke stappen je moet zetten om de meer subjectieve effecten van een behandeling goed te meten in een eenvoudig fictief onderzoek in de huisartspraktijk. Op deze manier hopen we enig inzicht te bieden in de begrippen die gebruikt worden om de waarde van dergelijke meetinstrumenten te beschrijven.

Probleemstelling

In een waarneemgroep wordt gedacht aan het opzetten van een onderzoek naar de waarde van terbinafine bij schimmelinfecties van handen of voeten. Eerder werden miconazolcreme en Whitfield-crème in een huisartspraktijk met elkaar vergeleken.[4] *Gezien de gunstige berichten over terbinafine lijkt een nieuw onderzoek op zijn plaats.*[5] *De groep wil daarbij niet alleen kijken naar 'genezen of niet', maar ook naar de hinder die patiënten van schimmels hebben. Er moet dus een instrument komen dat de welzijnseffecten van de behandeling meet.*

Het bepalen van de te meten aspecten

Maar hoe meet je welzijnseffecten? Is 'welzijn' in dit geval te vertalen als 'hinder' of 'ongemak'? Met andere woorden: welke aspecten van het hebben van een schimmelinfectie willen we meten? En vervolgens: welke eisen moeten we stellen aan het meetinstrument om het effect goed vast te kunnen stellen?

In de spreekkamer is een simpele vraag 'Heeft u er last van?' vaak voldoende om de patiënt te laten vertellen dat het jeukt, dat het er vies uitziet, dat hij zich schaamt, of niet durft te gaan zwemmen. De schoenen willen niet meer passen en de voeten doen af en toe pijn. Bij schimmelinfecties aan de nagels van de handen is werken soms niet prettig. De patiënt met een schimmelinfectie kent dus in de spreekkamer al een aantal aspecten toe aan 'last ervan hebben': somatisch, psychisch en sociaal ervaren hinder. Het gaat dus om de relevantie van de te meten aspecten. Bij sommige klachten is pijn een belangrijk

aspect, zodat de vragenlijst juist dat aspect goed moeten meten. Bij andere klachten is dat wellicht veel minder het geval.

In ons terbinafine-onderzoek vinden wij de algemene en psychische gezondheid niet zo belangrijk. Het lijkt immers niet zo waarschijnlijk dat een weliswaar hinderlijke maar niet ernstige kwaal bij de meeste patiënten een ingrijpende invloed op het bestaan zal hebben. Algemene schalen die juiste deze aspecten meten, zijn daarom niet zonder meer bruikbaar. Aspecten als zelfbeeld, hinder bij hobbies en werk en de subjectieve klachten, zoals specifieke pijn en jeuk door de aandoening, vinden we wel belangrijk. Ons meetinstrument moet dus ten minste deze aspecten kunnen meten. We willen daarbij niet alleen weten hoe iemand zich voelt, maar vooral of hij zijn gedrag door de aandoening aanpast. Bij het formuleren van de vragen zullen we daar dan ook op moeten letten.

De veel gebruikte generieke instrumenten meten alle deze verschillende aspecten, maar op de specifieke problemen van schimmelinfecties aan handen en voeten gaan ze natuurlijk niet in. Een specifiek 'schimmelinstrument' als aanvulling op een algemeen instrument zou dus zinvol kunnen zijn. De eerste vraag is dan of een dergelijk meetinstrument er al is.

Bij ons literatuuronderzoek naar 'health status' stuitten we op een onderzoek onder patiënten met schimmelnagels. *Lubeck et al.*[6] deden een onderzoek bij 680 deelnemers aan een Health Maintenance Organization (299 met en 381 zonder schimmelnagels). De vraag was of patiënten met onychomycose een lagere kwaliteit van leven hadden dan vergelijkbare mensen zonder schimmelaandoening.

De onderzoekers gebruikten daarvoor zes gestandaardiseerde algemene instrumenten: een algemene vraag ('hoe voelt u zich'), vier onderdelen uit de MOS SF-20 (pijn, psychische gezondheid, ervaren gezondheid, rolvervulling) en een instrument voor het meten van zelfwaardering/respect (Fleming Self-Esteem Scale). Omdat er nog geen specifieke 'schimmelinstrumenten' waren, ontwikkelden de onderzoekers '... scales that focused on patient-reported problems and concerns with physical appearance and physical activities. These items included specific limitations that might occur as a result of onychomycosis, such as standing on one's feet for long periods, or discomfort when using a computer keyboard or wordprocessor'.

Dit 'schimmelinstrument' lijkt op belangrijke punten aan onze wensen tegemoet te komen. Besloten wordt om het instrument en alle bijbehorende informatie op te vragen om de bruikbaarheid ervan voor ons onderzoek beter te kunnen beoordelen.

Algemene eisen te stellen aan een meetinstrument

Bij het kiezen of ontwerpen van een meetinstrument is het belangrijk je af te vragen *waarom* je iets wilt meten. Veelal kan dat uit de vraagstelling van het onderzoek worden afgeleid. Behalve omdat het in de mode is, zijn er drie redenen om effecten van een behandeling op de functionele toestand te meten:

- Je wilt beschikken over een *discriminatief* gegeven. In een beschrijvend onderzoek kan het nodig zijn, onderscheid te maken tussen mensen met en zonder beperkingen.
- Het meetinstrument heeft een *prognostische* functie: op grond van een score wil je iets zeggen over de toekomst. Bijvoorbeeld: voorspelt de vraag 'hoe fit voelt u zich nu?' de mortaliteit bij mannen tussen de 45-70 jaar?
- In een interventiestudie wil je ook iets zeggen over de subjectieve effecten: voelen patiënten zich ook beter (is er minder jeuk, pijn, ongemak?). Het meetinstrument wordt dan in *evaluatieve* zin gebruikt.[7]

Uiteraard kunnen meetinstrumenten aan verschillende doelen tegelijk trachten te voldoen.

Het instrument zal dus evaluatief gebruikt worden.

De eisen die aan een goed meetinstrument gesteld kunnen worden, hangen, behalve van het doel, ook af van de patiëntengroep waarin het gebruikt wordt. Een instrument dat vooral moet discrimineren tussen zieken en niet-zieken, heeft een andere inhoud dan een instrument dat vooral gevoelig moet zijn voor veranderingen. Zo is de vraag 'hoe moe voelt u zich?' ongeschikt voor het maken van onderscheid tussen zieke en niet-zieke ouderen, omdat moeheid in deze leeftijdgroep een te aspecifiek symptoom is. Maar de vraag is wel zinvol bij de evaluatie van de instelling van diabetespatiënten. Daarnaast heeft elk meetinstrument een signaal/ruis-verhouding. Alle discussies over de kwaliteit van instrumenten gaan over de beste manieren om het signaal, de ruis en de verhouding daartussen te meten.

Bij onze groep patiënten met schimmelinfecties is het stellen van de vraag 'hoeveel last heb je ervan?' voor en na de behandeling niet voldoende om het verschil in ervaren hinder precies te meten. Niet iedereen heeft immers jeuk of is bang te gaan zwemmen. Positieve antwoorden dekken dan verschillende ladingen en ook negatieve antwoorden behoeven niet te betekenen dat de patiënt nergens last van heeft. Misschien heeft hij niet aan dat aspect gedacht, of denkt hij er de eerste keer wel aan en de tweede keer niet.

Elk aspect wordt meestal met een aantal vragen of items gemeten; op die manier ontstaan schalen van bij elkaar horende vragen, waarmee wordt geprobeerd het signaal te versterken en de ruis te beperken. Dat de formulering van een vraag heel precies luistert, laat het volgende voorbeeld zien.

In een onderzoek met twee instrumenten bij HIV-geïnfecteerde mannen bleek er betrekkelijk weinig samenhang tussen twee schalen over pijn (r=0,66). Op het eerste gezicht leek dat vreemd, maar het bleek dat in het ene instrument met een enkele vraag gevraagd werd naar 'de duur van extreme pijn', terwijl de ander een enkele vraag bevatte over 'de ernst van pijn'. 'This indicates that brief Quality of Life scales, by necessity, can only address the most salient aspects of Quality of Life

and that complex phenomena such as pain cannot be comprehensively measured by a few items'.[8]

Om tot een aantal bruikbare items te komen wordt vaak een grote groep patiënten en/of deskundigen gevraagd welke items zij belangrijk vinden voor het doel en de populatie waarvoor het instrument wordt ontwikkeld. In verschillende stappen worden dan uit deze groslijst items verwijderd, eerst op grond van theoretische inzichten en later in proefonderzoeken op grond van statistische analyses.

Lubeck et al. *construeerden op deze manier drie schalen om het welzijn bij patiënten met schimmelnagels te meten: een schaal over 'physical appearance', bestaande uit acht items, een over problemen met dagelijkse activiteiten (tien items) en een over problemen met 'disease symptoms' van twaalf items. In de schaal 'dagelijkse problemen' zijn de volgende items opgenomen: 'discomfort from shoes, wearing any type of shoe, activities that require bare feet, hobby involving fingers e.g. knitting, hobbies requiring being on feet e.g. golf, performing daily work-related activities that expose nails, performing activities, such as typing, that use fingers, work activities that involve being on feet, e.g. sales person, construction, social activities, recreational activities'.*[6]

Alle vragen gaan specifiek over nagelproblemen. Omdat we niet alleen onderzoek doen naar nagelinfecties, zullen we de vragenlijst na vertaling moeten aanvullen. We moeten ervoor waken dat de vragenlijst dan niet te lang wordt.

Betrouwbaarheid

Na de opstelling en formulering van de vragen wordt het tijd ons te bekommeren over de testeigenschappen van de vragenlijst.

Aan het begrip betrouwbaarheid onderscheiden we een aantal aspecten. Zo is een meetinstrument betrouwbaar als het te meten aspect telkens op dezelfde wijze wordt vastgelegd. De metingen horen bijvoorbeeld onafhankelijk te zijn van het tijdstip van afname, de onderzoekssituatie, de interviewer, etc.

De test-hertest-correlatie is een eerste aspect van betrouwbaarheid en geeft aan in hoeverre eenzelfde persoon onder identieke omstandigheden ook dezelfde antwoorden op de vragen geeft. Als de 'objectieve' gezondheidstoestand van de patiënt niet verandert, behoort de score op de schaal ook niet te veranderen. Deze correlatie wordt veelal uitgedrukt in een Pearson's of produkt-moment-correlatiecoëfficiënt, waarbij geldt 'hoe hoger het getal, des te beter'.

Een tweede aspect van betrouwbaarheid is de de samenhang van de antwoorden op de bij elkaar horende vragen. De mate waarin bij elkaar horende items onderling samenhangen, wordt aangeduid met het begrip *interne consistentie*. Het scala aan statistische maten hiervoor is groot: Cronbach's alfa, item-rest-correlatie, betrouwbaarheidcoëfficiënt, etc.[9] De lezer van artikelen kan in eerste instantie volstaan met de wetenschap dat een hoog getal (>0,70) betekent dat het met de interne consistentie van de schaal wel goed zit. Bij de interpretatie van het getal moet wel het aantal items betrokken worden: een schaal met drie vragen en een Cronbach's alfa van 0,70 heeft een relatief hoge interne consistentie, terwijl een schaal met vijftien vragen en een alpha van 0,80 maar een matige interne consistentie heeft. Met een groter aantal vragen (een grotere steekproef uit alle theoretisch mogelijke items) heb je immers snel een betere verhouding tussen signaal en ruis.

De gevoeligheid van een instrument voor veranderingen of 'responsiviteit' wordt ook als een aspect van de betrouwbaarheid beschouwd, hoewel er ook veel voor te zeggen valt hier van een geheel aparte eigenschap van een meetinstrument te spreken. Bij gelijk blijvende klinische toestand moet er geen verandering in de score zijn, maar klinische veranderingen moeten zich vertalen in een andere score op het instrument. De responsiviteit is een maat voor het kleinste verschil (bijvoorbeeld in lichamelijke beperkingen voor en na behandeling), dat door het instrument gemeten kan worden.

Een algemeen gebruikte statistische maat voor responsiviteit is er nog niet. Een 'index of responsiveness' lijkt het meest aan te sluiten bij de klinische gedachtengang: de verhouding tussen het te voren gedefinieerde minimaal klinisch relevante verschil en de standaarddeviatie van patiënten die klinisch niet veranderen.[10] Om snel inzicht te krijgen in de samenhang tussen scores op een schaal en de klinische toestand, kunnen die het beste grafisch tegen elkaar worden uitgezet.[11]

Bij de responsiviteit is de breedte waarover het instrument meet – ook wel 'spreiding' genoemd – ook van belang: de plafond- en bodemeffecten moeten beperkt zijn. De score van patiënten moet kunnen verslechteren (er is niet zo snel een 'bodem'), maar ook bij een aanvankelijk hoge score moet deze nog kunnen verbeteren (het 'plafond' moet niet te snel bereikt zijn). Een goede spreiding ligt vaak besloten in de formulering van de vraag en de antwoordcategorieën. De vraag 'hoe snel kunt u op dit moment lopen?' met antwoordmogelijkheden 'heel snel' tot 'heel langzaam' geeft een betere spreiding dan dezelfde antwoordcategorieën bij de vraag 'kunt u in looppas lopen?'.

Bij de patiënten met schimmelinfecties willen we vooral de effecten van de toepassing van terbinafine meten. Het instrument moet dus gevoelig zijn voor veranderingen, zowel bij mensen die daarvan veel hinder ondervinden, als bij mensen met weinig hinder. In een proefonderzoek zullen we dus aan dit aspect voldoende aandacht moeten besteden. In dat proefonderzoek kan tegelijkertijd de interne consistentie en de test-hertest-correlatie van het vernieuwde instrument worden nagegaan.

Validiteit

Minstens zo belangrijk als de betrouwbaarheid van de meting is de vraag of we wel echt meten wat we willen meten. Hoe zit het met de validiteit van ons instrument?

Een meetinstrument is valide als het ook werkelijk meet wat gemeten moet worden: het 'signaal' dat wordt opgepikt, is ook het signaal waar het om gaat. Juist dit begrip zorgt voor de nodige verwarring, omdat niet iedereen hetzelfde onder begrippen als *content-, criterium-* of *constructvaliditeit* verstaat. Onze definitie van deze begrippen is maar een van de mogelijkheden, maar is in de praktijk goed bruikbaar.

De basis van het meten van validiteit is de *inhoudsvaliditeit* of *content validity*: komt de inhoud van de vragen overeen met de inhoud van het te meten begrip? Of het op het eerste gezicht zo'n beetje klopt heet ook wel *face validity*.

Als een meetinstrument wordt vergeleken met de echte werkelijkheid, wordt de *criteriumvaliditeit* onderzocht. Voor laboratoriumtests is het vinden van een gouden standaard ingewikkeld, maar voor vragenlijsten die welzijn meten, is het nog moeilijker en misschien zelfs onmogelijk. Bij instrumenten die de functionele toestand meten, lijkt het wat makkelijker om een gouden of 'vergulde' standaard te vinden. Het gaat dan meer om gedrag dan om het gevoel en gedrag is te obeserveren (kan de patiënt zich wel of niet aankleden, wandelen, etc).

Veel onderzoekers getroosten zich allerlei inspanningen om hun vragenlijst aan een gouden standaard te toetsen, zoals wordt geïllustreerd door de volgende twee voorbeelden.

In een onderzoek naar de validiteit van de SF-36 (een steeds vaker gebruikt instrument met verschillende schalen)[12] kiezen de auteurs als gouden standaard 'the first item on the SF-36, a single global health question... It is not common practice to use an item from a questionnaire to evaluate the criterion validity of that measure. The rationale for doing so is two fold. First, the item being used as the criterion variable is one that has been used as criterion variable in other studies. Second, the item contributes to only one dimension and does not contribute to the scales of the other seven dimensions.'

In een andere studie, bij mannen die een transurethrale prostatectomie (TUR) ondergingen, trachtten onderzoekers scores op de Nottingham Health Profile met een gouden standaard te vergelijken: 'here we report evidence of the criterion validity of the NHP by comparison of the Profile scores with other patient-recorded measures of pre and post operative status' (de mate van nycturie en nadruppelen).

'Omdat anderen het doen, doen we het ook maar' vinden wij geen goede redenering en vragen over nycturie en nadruppelen na een TUR hebben weinig kenmerken van een gouden standaard voor subjectieve gezondheid. Vaak bestaat een gouden standaard gewoon niet en is het streven naar vergelijking van het nieuwe meetinstrument daarmee verspilde moeite.[13 14]

We besluiten ons schimmelinstrument niet met een gouden standaard te vergelijken. Toch willen we iets meer over de validiteit van het instrument kunnen zeggen.

Bij gebrek aan een gouden standaard moet dus een andere vorm van validiteit gezocht worden. Zo kan de ene vragenlijst als referentie voor de andere worden gebruikt: een depressie-instrument dat al lang bestaat, kan als referentiekader dienen voor een nieuw instrument dat bijvoorbeeld voor een specifieke groep wordt ontwikkeld.[15] Deze vorm van validiteit heet *constructvaliditeit*. De score op de schaal van het nieuwe instrument moet dan samenhangen met de score op de vergelijkbare schaal van het oude instrument. Als de scores geheel hetzelfde zijn (een correlatie-coëfficiënt >0,75), levert de nieuwe schaal weliswaar weinig nieuws op, maar kan hij bijvoorbeeld voor patiënten minder belastend zijn om in te vullen. Deze vorm van constructvaliditeit heet wel *convergente* of *concurrente validiteit*.

Zo onderzochten *Van Marwijk et al.* bij patiënten tussen de 64 en 90 jaar de convergente validiteit van de nieuwe 'Geriatrische depressieschaal' (GDS) ten opzichte van de veel gebruikte 'Zung depressieschaal' (ZSDS): 'The correlation between ZSDS and GDS was high (0.74) in the present study. This indicates concurrent validity. Scale reliability of ZSDS and GDS was good (Cronbach's alpha 0.84 and 0.87). The GDS' yes-no format appears easier to fill in than the ZSDS'.[15]

Het omgekeerde, *divergente validiteit*, betekent dat er terecht geen verband is tussen de nieuw te testen schaal of instrument en een niet-verwante andere schaal of instrument.

Beide vormen van constructvaliditeit kunnen en moeten ook onderzocht worden door de schaal te vergelijken met bekende klinische condities: theoretisch moet een score dan wel of juist niet samenhangen met een stadium van een aandoening. Het toetsen van de NHP bij prostaathypertrofie is ons inziens dan ook meer constructvaliditeit. Maar door sommigen wordt juist criteriumvaliditeit gedefinieerd als de vergelijking van het meetinstrument met een theoretisch construct (de externe validiteit). Juist rond deze begrippen heerst dus verwarring.

Op grond van de formulering van de vragen in ons schimmelinstrument kan wel een uitspraak over de inhoudsvaliditeit gedaan worden en verder zou een hoge score op een 'hinderschaal' moeten samenhangen met de uitgebreidheid van de infectie. Het is immers te verwachten dat patiënten met een uitgebreide mycose op de hele voetzool meer hinder ondervinden dan patiënten met een klein plekje tussen twee tenen. We kunnen dus ook een uitspraak doen over de constructvaliditeit. We gaan er dan maar van uit dat we de 'uitgebreidheid van de infectie' goed en betrouwbaar kunnen vastleggen, bijvoorbeeld door de oppervlakte van de aandoening te meten.

Het op deze wijze valideren van een instrument doet een beetje denken aan het verhaal waarin de Baron van Münchausen zichzelf (en zijn paard) aan zijn eigen haren uit een moeras omhoogtrok. Door een meetinstrument keer op keer en in verschillende situaties te onderzoeken, kan die lijst steeds meer valide worden. Het lijken op het eerste gezicht dubieuze trucs, maar toch is het bij het ontbreken van een gouden standaard *second best*. In de spreekkamer hebben we immers ook geen

gouden standaard, als we vragen hoe het met iemand gaat.

Beschouwing

Bij het meten van de 'effecten' in ons hypothetisch onderzoek over de behandeling van schimmelinfecties stuitten we op een aantal vragen die bij elk onderzoek naar het effect van een behandeling belangrijk zijn en om een antwoord vragen: welke aspecten zijn zinvol om te meten, gaat het om subjectieve beleving of willen we meer de functionele toestand meten?

Het antwoord op deze vragen moet vooral op inhoudelijke argumenten gebaseerd zijn. Dat de onderzoekers beter over de inhoud van hun instrumenten (inhoudsvaliditeit) moeten nadenken, werd recent aangetoond door *Gill & Feinstein*.[16] Bij meer dan de helft van de 75 at random getrokken artikelen waarbij 'quality of life' gemeten werd, maakten de onderzoekers niet duidelijk op grond van welke ideeën ze de meetinstrumenten hadden uitgekozen en waarom ze juist die aspecten wilden weten. Lang niet altijd is het ons inziens nodig de kwaliteit van leven in onderzoek gedetailleerd te weten. Het is maar zeer de vraag of de 'kwaliteit' van leven van patiënten met schimmelnagels wel zo interessant is.

Zowel bij het opzetten van eigen als het beoordelen van andermans onderzoek is het nuttig om eerst op de stoel van de huisarts en de patiënt te gaan zitten: is er wel een klinisch relevant terrein in kaart gebracht en wordt dit op een voor de praktijk relevante manier onderzocht? Is het antwoord op deze vraag negatief, dan is het artikel voor de praktijk niet interessant.

Pas bij een positief antwoord zijn de methodologische aspecten belangrijk:
- met welk doel worden effecten gemeten;
- in welke populatie vindt het onderzoek plaats;
- hoe betrouwbaar en valide is het te gebruiken instrument?

Indien bestaande instrumenten, die veelal uitgebreid op betrouwbaarheid en validiteit zijn onderzocht, geschikt zijn, genieten deze de voorkeur, omdat de gegevens in verschillende onderzoeken dan vergelijkbaar zijn. Recent werd daarom aanbevolen voorlopig een betrekkelijk simpel meetinstrument als de COOP/WONCA-kaarten als 'meeloper' te gebruiken.[17] Een vooronderzoek naar de betrouwbaarheid en validiteit kan dan vaak achterwege blijven en het is mogelijk om de resultaten te vergelijken met andere onderzoeken.

Omdat de bestaande instrumenten vaak niet volledig aan de wensen van de onderzoeker tegemoet komen, is aanvulling met – liefst zo weinig mogelijk – nieuwe schalen gebruikelijk. Het kunnen overleggen van gegevens over betrouwbaarheid en validiteit van deze nieuwe schalen op basis van een proefonderzoek verhoogt de waarde van de resultaten van het hoofdonderzoek.

Om onze belofte waar te maken om voor niet ingewijde lezer helderheid te scheppen geven we tot slot een aantal vragen voor het beoordelen van onderzoek waarin effecten zijn gemeten:
- Welke welzijsaspecten zijn gemeten: subjectieve klachten, functionele toestand, of kwaliteit van leven?
- Is het doel van het meten van welzijnsaspecten (discriminatie, voorspelling of eveluatie) duidelijk aangegeven?
- Past het gekozen instrument bij het doel van de meting?
- Wordt de keuze door de auteurs gemotiveerd?
- Welke verschillende dimensies bevat het gebruikte instrument?
- Zijn deze dimensies relevant voor de patiënt en/of voor de huisarts?
- Worden gegevens vermeld over de betrouwbaarheid en validiteit van het gekozen instrument?

Literatuur

1 Guyatt GH, Kirskner B, Jaescke R. A methodological framework for health status measurement; clarity or oversimplification? J Clin Epidemiol 1992; 45; 1353-5.

2 De Neeling JND. Quality of life; het onderzoek naar welzijnseffecten van medische behandelingen. Utrecht: Bunge 1991.

3 König-Zahn C, Furer JW, Tax B. Het meten van de gezondheidstoestand; algemene gezondheid. Assen: Van Gorcum, 1993.

4 Grooten JAM, Meijman FJ. De behandeling van dermatomycosen van tenen/voeten en liezen in de huisartspraktijk. Een vergelijkend onderzoek naar de effectiviteit van miconazol en Whitfield-crème. Huisarts Wet 1992; 35; 26-8.

5 Goodfield MJD, Andrew L, Evans EGV. Short term treatment of dermatophyte onychomycosis with terbinafine. BMJ 1992; 304: 1151-4.

6 Lubeck DP, Patrick DL, McNulty P, et al. Quality of life of persons with onychomycosis. Qual Life Res 1993; 2: 341-8.

7 Guyatt GH, Kirshner B, Jaescke R. Measuring health status: what are the necessary measurement properties? J Clin Epidemiol 1992; 45: 1341-5.

8 Burgess A, Dayer M, Catalan J, et al. The reliability and validity of two HIV-specific health-related quality-of-life measures: a preliminary analysis. AIDS 1993; 7: 1001-8.

9 Feinstein AR. Clinimetrics. New Haven, Conniticut: Yale University Press, 1987.

10 Grootenhuis PA, Snoek FJ, Heine RJ, Bouter LM. Development of a Type 2 Diabetes Symptom Checklist; a measure of symtom severity. Diabetes Med 1994; 11: 253-261.

11 Schuling J, Greidanus J, Meyboom-de Jong B. Measuring functional status of stroke patients with the Sickness Impact Profile. Disabil Rehabil 1993; 15: 19-23.

12 Jenkinson C, Wright L, Coulter A. Criterion validity and reliability of the SF-36 in a population sample. Qual Life Res 1994; 4:7-12.

13 Jacobs HN. Health status measurement in family medicine research [Dissertatie]. Utrecht: Universiteit Utrecht, 1993.

14 Patrick DL, Erickson P. Health status and health policy; allocating resources to helath care. Oxford: Oxford University Press, 1993: 198-202.

15 Marwijk H, Arnold I, Bonnema J, Kaptein A. Self-report depression scales for elderly patients in primary care; a preliminary study. Fam Practice 1993; 10; 63-5.

16 Gill TM, Feinstein AR. A critical appraisal of the quality of quality-of-life measurements. JAMA 1994; 272: 619-26.

17 Essink-Bot ML. De werkgroep Onderzoek Gezondheidstoestandmeting; standaardisatie onderzoek naar met gezondheid samenhangende kwaliteit van leven. Ned Tijdschr Geneeskd 1994; 138: 1484-6.

■