De keuze van een vragenlijst

Methodologische en praktische overwegingen

C. KÖNIG-ZAHN
J.W. FURER

König-Zahn C, Furer JW. De keuze van een vragenlijst. Methodologische en praktische overwegingen. Huisarts Wet 1995; 38(3): 110-6, 128.

Samenvatting Het meten van de ervaren gezondheidstoestand krijgt al enige jaren grote aandacht. Daarbij zijn veel vragenlijsten beschikbaar gekomen, die niet allemaal voldoen aan de hoogste normen. Om tot een juiste keuze te komen is inzicht gewenst in de diversiteit aan instrumenten en in de wetenschappelijke en praktische eisen die aan een instrument gesteld kunnen worden. Bij de keuze van een instrument zal de inhoudelijke vraagstelling van het onderzoek de belangrijkste afweging zijn. Is de vraagstelling sterk op een aandoening toegesneden, dan zal zeker een ziektespecifiek instrument gezocht moeten worden, terwijl een breed, generiek instrument ook in zulke onderzoeken van nut is om de resultaten in een algemeen kader te kunnen plaatsen. De wetenschappelijke eisen (validiteit, betrouwbaarheid, gevoeligheid voor verandering) worden kort besproken. Praktische overwegingen kunnen stringente eisen opleggen waaraan veel instrumenten niet zullen voldoen.

Vakgroep Huisartsgeneeskunde, Sociale Geneeskunde en Verpleeghuisgeneeskunde, Katholieke Universiteit Nijmegen, Postbus 9101, 6500 HB Nijmegen.
C. König-Zahn, arts-epidemioloog, Drs. J.W. Furer, psycholoog.
Correspondentie: C. König-Zahn.

Inleiding

Op het terrein van de gezondheidsmeting is de laatste decennia een overstelpend aantal vragenlijsten verschenen: *Spilker et al.* verzamelden in hun bibliografie, die beperkt bleef tot de Engelstalige onderzoeksliteratuur, meer dan driehonderd instrumenten.[1] Deze explosieve groei maakt duidelijk dat bij het bestuderen van de effecten van medische interventies een ruimer gezondheidsconcept ingang heeft gevonden, waarin de ervaren gezondheid en de functionele toestand van patiënten een aanvulling vormen op de traditionele biomedische en andere klinische effectmaten.

De meerderheid van de door *Spilker et al.* verzamelde vragenlijsten was ontwikkeld voor klinische problemen, voornamelijk als ziektespecifieke effectmaten in therapeutische trials. Voor een deel ging het daarbij om 'eendagsvliegen': vragenlijsten die in het kader van één bepaald onderzoek zijn toegesneden op specifieke vraagstellingen, en daarna nooit meer worden gebruikt. Aan de meetkwaliteit van dit soort instrumenten wordt meestal weinig aandacht besteed – een blijk van onderschatting van de methodologische eisen die men aan een vragenlijst moet stellen.

In deze bijdrage geven wij een overzicht van de verschillende aspecten die van belang zijn bij de keuze van een vragenlijst. Aan de orde komen de volgende onderwerpen:

- Het onderscheid tussen generieke en ziektespecifieke vragenlijsten, en hun voor- en nadelen.
- De methodologische eisen waaraan vragenlijsten moeten voldoen:
  - validiteit;
  - betrouwbaarheid;
  - gevoeligheid voor verandering.
- De praktische overwegingen die bij de keuze van een vragenlijst een rol spelen.

Literatuur

De geraadpleegde literatuur is verzameld in het kader van het project 'Het meten van de gezondheidstoestand: beschrijving en evaluatie van vragenlijsten'. Enkele jaren geleden is begonnen met de opbouw van dit literatuurbestand door het raadplegen van Medline en Psyclit. Daarbij werd vooral gezocht naar publikaties over de ontwikkeling van instrumenten en hun meeteigenschappen. Dit bestand bracht ons vervolgens op het spoor van diverse boeken over het meten van de ervaren gezondheidstoestand.

Voor deze bijdrage hebben wij vooral de goed toegankelijke handboeken en enkele frequent aangehaalde artikelen gebruikt. Voor de vragenlijsten die hier als voorbeelden zijn genoemd, hebben wij telkens een oorspronkelijk en – als dat beschikbaar was – een Nederlands artikel in de literatuurlijst opgenomen.

Generiek of ziektespecifiek: algemeen of op maat

Met vragenlijsten kan informatie over diverse aspecten van de gezondheidstoestand worden verworven. Deze informatie dient niet als vervanging van de diagnostische methoden van de arts, maar zijn bedoeld om inzicht te krijgen in de aspecten die daarbij over het algemeen niet tot hun recht komen. De gevolgen van ziekte op welbevinden en functioneren staan centraal; deze terreinen worden aangeduid met de termen 'ervaren gezondheid' en 'functionele toestand', door ons in het vervolg *ervaren gezondheidstoestand* genoemd. Gezondheidstoestand is een breed begrip en bevat volgens de WHO-gezondheidsdefinitie de dimensies lichamelijke, psychische en sociale gezondheid.

Generiek versus ziektespecifiek

Vele vragenlijsten besteden aandacht aan al deze dimensies, zoals de SIP (Sickness Impact Profile),[2,3] de NHP (Nottingham Health Profile),[4-7] de COOP/WONCA-kaarten[7-9] en de recent in Nederland geïntroduceerde SF-36.[10-12] Deze instrumenten zijn gericht op de gevolgen van een breed spectrum van gezondheidsstoornissen en aandoeningen en worden *generieke vragenlijsten* genoemd. In principe zijn zij geschikt voor alle categorieën patiënten, en ook voor niet-patiënten.

Met deze generieke instrumenten zijn vergelijkingen tussen verschillende aandoeningen en interventies mogelijk, en zij worden dan ook vaak in MTA-onderzoek toegepast. Hun nadeel is dat zij over het algemeen vrij lang zijn, en dat zij soms onvoldoende op specifieke aandoeningen zijn toegesneden. Zij besteden weinig aandacht aan voor een specifieke aandoening zeer relevante gevolgen voor bijvoorbeeld het functioneren, terwijl ook een reeks voor deze aandoening irrelevante vragen beantwoord moet worden. Daardoor zijn deze instrumenten soms minder gevoelig voor veranderingen van de aan een bepaalde aandoening gerelateerde functionele toestand.

Met name in evaluatie-onderzoek worden dan ook vaak *ziektespecifieke vragenlijsten* gebruikt, die toegesneden zijn op een bepaalde aandoening en/of een bepaalde behandeling. Het voordeel zal duidelijk zijn, het nadeel is dat geen vergelijking tussen verschillende aandoeningen meer mogelijk is. Bekende ziektespecifieke instrumenten zijn de AIMS (Arthritis Impact Measurement Scale[13]), de DHP-1 (Diabetes Health Profile,[14] vertaald door B. Meyboom-de Jong, Vakgroep Huisartsgeneeskunde, Groningen), de CRDQ (Chronic Respiratory Disease Questionnaire,[15] vertaald door MPMH Rutten-van Mölken, Vakgroep Economie van de Gezondheidszorg, Maastricht), en de CES-D (Center for Epidemiologic Studies Depression Scale,[16] vertaald door A.T.F. Beekman, D. Deeg en W. van Tilburg, Vakgroep Psychiatrie Vrije Universiteit Amsterdam, en J. van Limbeek en L. Wouters, GG en GD Amsterdam).

Om het verschil in benadering te illustreren hebben wij in het *kader* op deze pagina de vragen naar het lichamelijke functioneren uit de generieke SF-36 en de ziektespecifieke CRDQ naast elkaar gezet.

Voorbeeld van een generieke en een ziektespecifieke vragenlijst – fysiek functioneren in de SF-36 en de CRDQ

SF-36

Fysiek functioneren

De volgende vragen gaan over dagelijkse bezigheden. Wordt u door uw gezondheid op dit moment beperkt bij deze bezigheden? Zo ja, in welke mate?*

- forse inspanning, zoals hardlopen, zware voorwerpen tillen, inspannend sporten
- matige inspanning, zoals het verplaatsen van een tafel, stofzuigen, fietsen
- tillen of boodschappen doen
- een paar trappen oplopen
- één trap oplopen
- buigen, knielen, of bukken
- meer dan een kilometer lopen
- een halve kilometer lopen
- honderd meter lopen
- uzelf wassen of aankleden

CRDQ

Fysieke activiteit

Kunt u beschrijven hoe kortademig u zich de afgelopen 2 weken heeft gevoeld bij het uitvoeren van ieder van de 5, voor u meest belangrijke aktiviteiten?†

- boos of overstuur zijn
- baden of douchen
- voorover buigen
- dragen, zoals boodschappen
- aankleden
- eten
- een stukje wandelen
- huishoudelijk werk doen
- haast maken
- plat liggen
- bed opmaken
- de vloer dweilen of schrobben
- meubiliar verplaatsen
- met kinderen of kleinkinderen spelen
- sport en spel beoefenen
- boven het hoofd reiken
- hollen, zoals om de bus te halen
- winkelen
- praten
- stofzuigen
- in en rondom het huis lopen
- een helling oplopen
- een trap oplopen
- met anderen op vlak terrein lopen
- maaltijden bereiden
- in slaap proberen te vallen

* De antwoordmogelijkheden zijn: ja, ernstig beperkt – ja, een beetje beperkt – nee, helemaal niet beperkt.

† Voordat de ondervraagde deze vraag beantwoordt moet hij eerst uit deze lijst de 5 voor hem belangrijkste activiteiten kiezen. Voor elke van deze 5 activiteiten geeft hij vervolgens op een zevenpuntsschaal aan hoe kortademig hij daarbij is geweest.

Structuur

Generieke instrumenten bestrijken in principe verscheidene dimensies van gezondheid. Expliciet worden diverse deelconcepten onderscheiden en geoperationaliseerd als afzonderlijke schalen, die dan deel uitmaken van één meerschalig instrument. Ook in de ziektespecifieke instrumenten zijn meestal verschillende gezondheidsconcepten geoperationaliseerd en zij bevatten over het algemeen dus eveneens verscheidene schalen.

Onder een schaal wordt een verzameling van items verstaan die alle één aspect van de ervaren gezondheidstoestand meten. Voordelen van het gebruik van een meer-item-schaal zijn een hogere betrouwbaarheid – de schaalscore is minder onderhevig aan toevalsfluctuaties – en een grotere precisie. Er bestaan verschillende technieken om tot een schaal te komen; deze zijn gebaseerd op verschillende schaalmodellen. Bekende voorbeelden zijn de Guttman-, de Thurstone- en de

Likertschaal. Een uitvoerige beschrijving van gangbare schaaltechnieken is te vinden bij *Swanborn.*[17]

Uit de antwoordscores van de items wordt een uiteindelijke schaalscore berekend. De afzonderlijke schaalscores worden over het algemeen naast elkaar gepresenteerd, en de meerschalige instrumenten worden daarom ook vaak aangeduid als *profiel*. Meerschalige vragenlijsten, waarvan de schaalscores tot één totale score worden samengevat, worden veelal *index* genoemd.

Het aantal items per schaal kan sterk uiteenlopen. Zo zijn in de COOP/WONCA-kaarten vijf concepten door middel van telkens één item geoperationaliseerd, en de bij elke vraag gegeven vijf antwoordmogelijkheden vormen een ordinale vijfpuntsschaal. De SIP heeft twaalf schalen met per schaal 7-23 items. Voordeel van de COOP/WONCA-kaarten is dat men met een korte vragenlijst een breed spectrum van de ervaren gezondheidstoestand in kaart kan brengen. Nadelen zijn een geringere betrouwbaarheid en een geringer vermogen tot differentiatie.

Er zijn verder nog vragenlijsten die gekarakteriseerd kunnen worden met het predikaat *thermometer*, een term die *Dohrenwend & Dohrenwend* hebben geïntroduceerd.[18] Aan zulke vragenlijsten ligt geen eenduidig concept ten grondslag. De vragen slaan op uiteenlopende gezondheidsconcepten, maar worden toch samengevat tot één schaalscore, die dan als het ware bij een 'verhoging' aangeeft dat er iets aan de hand is, maar niet wat dat precies is. Een van de oudste vragenlijsten, de Cornell Medical Index, is een voorbeeld van zo'n thermometer.

Methodologische eisen – algemeen

De gangbare beoordelingscriteria voor de meeteigenschappen (psychometrische eigenschappen) van vragenlijsten zijn de validiteit en de betrouwbaarheid. Daarnaast is bij de meting van de ervaren gezondheidstoestand, met name als deze als effectmaat bedoeld is, de 'gevoeligheid voor verandering' van belang. Deze gevoeligheid, in het Engels 'responsiveness' of 'sensitivity to change' genoemd en omschreven als 'a questionnaire's ability to detect clinically important changes in patient status',[19] is zonder twijfel een belangrijk kenmerk. Controversieel is of de gevoeligheid, naast validiteit en betrouwbaarheid, te beschouwen is als een zelfstandig concept, en zo ja, hoe zij te meten is.

Ondanks of misschien juist als gevolg van een lange traditie van het validiteits- en betrouwbaarheidsonderzoek, voornamelijk in de psychologie en sociologie, is de in de onderzoeksliteratuur voorkomende terminologie enigszins verwarrend. Daarom zullen deze termen en hun achtergronden kort worden uitgelegd. Hierbij worden de aanbevelingen van de American Psychological Association gevolgd.[20] De daar gehanteerde definities zijn met name in de Verenigde Staten ook bij vele ontwikkelaars van vragenlijsten ingeburgerd geraakt voor de meting van de ervaren gezondheidstoestand.

Validiteit

Gezondheid en ziekte zijn complexe fenomenen die zich aan een rechtstreekse observatie onttrekken. Daarom is een vragenlijst die beoogt de ervaren gezondheidstoestand te meten, niet meer dan een indicator van een theoretisch gezondheidsbegrip.* Dit theoretische begrip is ingebed in een netwerk van relaties met andere gezondheidsbenvloedende factoren en door gezondheid beïnvloede begrippen. Over deze relaties tussen het begrip 'ervaren gezondheidstoestand' (zoals door het meetinstrument geïndiceerd) en andere factoren (bijvoorbeeld leeftijd of de aanwezigheid van een chronische aandoening) kunnen specifieke hypothesen worden opgesteld. Om als valide beschouwd te kunnen worden moet deze indicator van de gezondheidstoestand zich zo gedragen als op basis van een algemeen aanvaarde theorie wordt verondersteld.

Het zal duidelijk zijn dat het formuleren van (een reeks van) hypothesen vergt dat duidelijkheid bestaat over de conceptuele achtergrond van het meetinstrument. Met dit proces van hypothesen-toetsend onderzoek bouwt een vragenlijst *begripsvaliditeit* ('construct validity') op. Een van de gebruikelijke criteria in dit proces is dat van een hoge correlatie tussen de vragenlijst en gelijkwaardige indicatoren van het betreffende gezondheidsbegrip – convergente validiteit ('convergent validity') genoemd – en van een lage correlatie tussen de vragenlijst en indicatoren van andere (gezondheids)begrippen – divergente of discriminante validiteit ('discriminant validity') genoemd.

Wij geven de voorkeur aan het gebruik van divergente in plaats van discriminante validiteit, omdat discriminante validiteit in het Nederlands ook gebruikt wordt om aan te geven dat een instrument in staat is te discrimineren tussen van elkaar verschillende groepen respondenten (bijvoorbeeld tussen patiënten en gezonden). Dit onderscheidend vermogen van een instrument wordt in het Engels aangeduid met de term 'known-group validity' en is een eerste vereiste voor gevoeligheid.

Een zeer stringente manier om begripsvaliditeit aan te tonen is de 'multitrait-multimethod'-aanpak (MTMM-aanpak) volgens *Campbell & Fiske.*[21,22] Hierbij worden de correlaties tussen diverse schaalscores systematisch op het te verwachten patroon gecontroleerd. Convergente validiteit is aanwezig, wanneer de correlaties tussen verschillende schalen die geacht worden hetzelfde kenmerk te meten – bijvoorbeeld overeenkomstige of verwante schalen van verschillende vragenlijsten – hoog zijn. Een indicatie voor divergente validiteit is een lage correlatie tussen verschillende kenmerken binnen één vragen-

* Volgens *De Groot* is een theorie 'een systeem van logisch samenhangende beweringen, opvattingen en begrippen betreffende een werkelijkheidsgebied, die zo zijn geformuleerd dat het mogelijk is er een toetsbare hypothese uit af te leiden'. De meeste hypothesen gaan over de relatie tussen begrippen. Al naar gelang hun plaats in een veronderstelde causale keten worden begrippen ook met de term *factor* aangeduid. Begrippen kunnen op een continuüm van abstract naar concreet worden geplaatst. De begrippen aan de abstracte kant worden concepten of hypothetische begrippen genoemd, die aan de concrete kant empirische of operationele begrippen, variabelen of indicatoren.[29,30]

lijst én een lage correlatie tussen verschillende kenmerken gemeten met verschillende vragenlijsten (*tabel*). Zo zouden de scores van de schaal 'lichamelijk functioneren' van bijvoorbeeld de COOP/WONCA-kaarten, de NHP en de SIP hoog met elkaar moeten correleren (convergente validiteit): zij beogen immers hetzelfde kenmerk te meten; anderzijds zouden de correlaties van deze schaalscores met bijvoorbeeld de schaalscores 'sociaal functioneren' afkomstig van dezelfde dan wel een andere lijst, laag moeten zijn (divergente validiteit).

Het opbouwen van begripsvaliditeit kan – gezien de vele mogelijke hypothesen en de verscheidenheid aan onderzoeksopzetten – via verschillende wegen lopen en, naarmate de theorie omvangrijker is, meer onderzoek vergen.

Een belangrijke vraag is in welke mate de vragenlijst het gezondheidsbegrip representeert. De mate van representatie wordt de *inhoudsvaliditeit* ('content validity') genoemd. Dit validiteitsaspect is niet met statistische of andere kwantitatieve technieken te benaderen. Het gaat hier om de inschatting of het theoretische en meestal vrij abstracte begrip inhoudelijk adequaat is ingevuld en of de items dat vervolgens goed weerspiegelen. De inhoudsvaliditeit kan alleen beoordeeld worden wanneer de conceptuele achtergronden van het meetinstrument duidelijk zijn beschreven. Verder is belangrijk dat de herkomst van de items toegelicht wordt: hoe zijn de items verkregen (bijvoorbeeld op basis van adviezen van artsen, door ondervraging van patiënten, door literatuurstudie), en hoe is men tot de keuze van de uiteindelijke items gekomen (bijvoorbeeld op inhoudelijke en/of statistische gronden).

Indien voor het te meten gezondheidsbegrip een criterium bestaat of een 'gouden standaard' aanvaard wordt, is een van de eerste onderzoeksstappen het bepalen van de relatie tussen de vragenlijst en het criterium. Deze vorm van validiteit wordt *criterium-validiteit* of *criterium-gerelateerde validiteit* ('criterion-related validity') genoemd. Daarbij kan het om een gelijktijdige meting van beide gaan ('concurrent validity') of om een aan de criteriumvaststelling voorafgaande bepaling van de score op de vragenlijst ('predictive validity'). Het criterium is een door een andere methode verkregen waarneming zonder fouten of, als een dergelijke methode niet voorhanden is, een algemeen aanvaard criterium, de 'gouden standaard'.

Op de terreinen van de ervaren gezondheid en de functionele toestand bestaat noch een foutloos noch een bijna foutloos, algemeen aanvaard criterium dat men als een 'gouden standaard' zou kunnen laten gelden. De correlaties met verwante instrumenten zijn niet te interpreteren als criteriumvaliditeit; ook ingeburgerde goede vragenlijsten meten niet zo foutloos dat zij als 'gouden standaard' kunnen dienen. Zoals eerder bij de beschrijving van de MTMM-aanpak beschreven is, zijn deze verbanden te beschouwen als indicaties voor de begripsvaliditeit van de (nieuwe) vragenlijst.

Naast de drie belangrijke 'typen' validiteit (begrips-, inhouds- en criteriumvaliditeit) bestaan er nog enkele min of meer verwarrende termen. Zo worden 'trait validity' en 'factorial validity' als synoniemen voor begripsvaliditeit gebruikt, en 'empirical validity' en 'statistical validity' als synoniemen voor criteriumvaliditeit. De term 'face-validity' slaat op de inzichtelijkheid en de plausibiliteit van de vragen, en is volgens de American Psychological Association niet als een type validiteit te beschouwen, maar als een facet van de gebruiksvriendelijkheid.

Betrouwbaarheid

Terwijl het het bij de validiteit om de inhoud gaat, om wát er gemeten wordt, heeft de betrouwbaarheid betrekking op de reproduceerbaarheid van de meting. Van een vragenlijst mag verwacht worden dat – bij een gelijke toestand van de ondervraagde – bij herhaalde meting dezelfde score verkregen wordt. De mate van overeenstemming tussen de antwoorden op verschillende tijdstippen wordt *test-hertest-betrouwbaarheid* ('test-retest reliability') genoemd. Maten voor de betrouwbaarheid zijn overeenstemming (bijvoorbeeld Cohen's kappa) of correlaties (bijvoorbeeld intra-class correlatiecoëfficiënt).

Een ander aspect van de betrouwbaarheid betreft de vragenlijst op zichzelf en gaat over de *homogeniteit* van (een reeks van) vragen, de *interne consistentie* van een schaal. Daar een vraag niet door elke persoon op elk moment gelijk begrepen wordt, ontstaat door toevalsfluctuatie een meetfout. Om deze meetfout terug te dringen, worden zoveel vragen (die zich uiteraard op hetzelfde achterliggende concept richten) aan de schaal toegevoegd dat de gewenste mate van homogeniteit bereikt wordt en de meetfout acceptabel blijft. Vuistregels hierbij zijn dat men streeft

Tabel *MTMM-correlatiematrix voor twee denkbeeldige meerschalige vragenlijsten*

	Aa	Ab	Ac	Ba	Bb
Aa	bet*				
Ab	laag†	bet*			
Ac	laag†	laag†	bet*		
Ba	hoog§	laag‡	laag‡	bet*	
Bb	laag‡	hoog§	laag‡	laag†	bet*

A en B staat voor twee verschillende vragenlijsten ('methods')
a b en c staat voor verschillende gezondheidsaspecten ('traits')
* test-hertest betrouwbaarheid (zie paragraaf betrouwbaarheid)
† 'heterotrait-monomethode' correlaties
‡ 'heterotrait-heteromethode' correlaties
§ 'monotrait-heteromethode' correlaties

naar een coëfficiënt voor homogeniteit van 0,80, hetgeen haalbaar geacht wordt met tien niet-dichotome of twintig dichotome items.[23] De coëfficiënt voor niet-dichotome items is de alfa van Cronbach, de coëfficiënt voor dichotome items is de Kuder-Richardson-formule nummer 20 (KR20).

Gevoeligheid voor verandering

Terwijl voor validiteit en betrouwbaarheid (min of meer) gestandaardiseerde terminologie, methoden en statistiek beschikbaar zijn, is daarvan nog geen sprake voor de gevoeligheid voor verandering ('responsiveness'). De vraag naar het vermogen van een vragenlijst om 'zelfs kleine, klinisch relevante veranderingen in de tijd waar te nemen', lijkt een logische wens uit de praktijk naar de inzetbaarheid als effectmaat van de diverse vragenlijsten. Het lijkt hierbij te gaan om een combinatie van validiteit en betrouwbaarheid: de vragenlijst moet betrekking hebben op *klinisch relevante* veranderingen in de ervaren gezondheidstoestand (validiteit) en zelfs *kleine* veranderingen kunnen onderscheiden (weinig ruis bevatten: betrouwbaarheid).

Naarmate de ruis (error, fout) in de meting toeneemt, neemt de gevoeligheid voor verandering af. Als ideaaltypische procedure om de gevoeligheid vast te stellen, stellen *Jaeschke & Guyatt* voor twee onafhankelijke onderzoeken te doen.[24] In het ene moet de ruis worden opgespoord door bij personen met een stabiele (maar onderling uiteenlopende) ervaren gezondheidstoestand herhaaldelijk te meten. In het andere moet bij een behandeling waarvan het klinisch effect bekend is, worden nagegaan of het meetinstrument het effect van die behandeling juist weergeeft. De ratio tussen de verandering bij de behandelde groep en de ruis bij de stabiele groep zou dan een schatting mogelijk kunnen maken van de gevoeligheid voor verandering van een instrument.

In de reële onderzoeksituatie van de clinical trial suggereren *Guyatt et al.* als gevoeligheidsmaat de ratio tussen enerzijds de verschuiving bij de behandelde groep en anderzijds de schatter voor ruis verkregen uit herhaalde metingen van de controlegroep.[25]

Mogelijk ten overvloede wijzen wij erop dat bij het bepalen van de effectiviteit van een behandeling of interventie niet alleen de gevoeligheid voor verandering van het meetinstrument van belang is. Zo speelt uiteraard het onderzoeksdesign een rol (is er gerandomiseerd, hoeveel behandelingscondities zijn er, welke sterkten hebben die condities, welke zijn de groepsgrootten, enz.). Ook de wijzen van statistische bewerking en toetsing zijn van belang (verschilscores kennen een verminderde betrouwbaarheid, past de frequentieverdeling van de score bij de statistische toets, enz.).[26]

De bron van de gevoeligheid voor verandering zetelt uiteraard in (de gevoeligheid van) de vragen. Bijna onveranderlijke toestanden moeten als onderwerp vermeden worden. De antwoorden bij eerste meting op een vraag als 'Hebt u weleens... gedaan/gehad?' zullen bij tweede meting waarschijnlijk weinig verandering vertonen. Als voorbeeld hiervan kan de SIP-vraag 'Ik heb geprobeerd een eind aan mijn leven te maken' gelden.[27] Ook gebruik van vragen over gezondheidstoestanden die vrijwel alleen eenzijdig zullen veranderen (heel gezond kan nauwelijks beter, heel slecht kan nauwelijks zieker), moet vooraf goed worden afgewogen. Een voorbeeld van een te zware vraag die door bijna niemand met ja wordt beantwoord, is de SIP-vraag 'Ik eet niet zelfstandig, maar moet gevoed worden'.[27]

De gevoeligheid kan vergroot worden door in de antwoorden veel variatiemogelijkheden aan te brengen, bijvoorbeeld met een 'visual analog scale' of met vragen van het Likert-type met bij voorkeur vijf of meer (maar wel betekenisvolle) antwoordcategorieën.[24] Ook wordt bij de vervolgmeting wel gewerkt met de antwoorden van de vorige afname, opdat een verandering beter ingeschat kan worden door de behandelde.[24]

Ziektespecifieke meetinstrumenten kunnen goed inspelen op de typische, veranderlijke symptomen en het verloop van een aandoening. Zij kunnen tevens nauw aansluiten op de medische denk- en werkwijze, en daarmee de klinisch relevante veranderingen bijna inherent weergeven. De aansluiting op de interventie kan eveneens scherp worden afgesteld.

Daar staat tegenover dat deze vragenlijsten per definitie toegespitst zijn, en dat geen vergelijking mogelijk is tussen interventies en tussen aandoeningen. Zulke meetinstrumenten staan dus geïsoleerd, ze missen de aansluiting op een algemeen kader. Daar er vele aandoeningen, behandelingen en doelgroepen te bestuderen zijn, zullen ontelbare specifieke instrumenten nodig zijn, die elk zorgvuldig ontworpen en getest moeten worden.[28]

Over de betrouwbaarheid en validiteit van generieke vragenlijsten is daarentegen veel bekend. Weliswaar zijn deze instrumenten niet zo gevoelig voor de specifieke, typische elementen van een bepaalde aandoening en behandeling, maar bij een interventie die de ervaren gezondheid en/of functionele toestand in sterke mate bevordert – bij welke specifieke aandoening dan ook – zijn zij niettemin goed bruikbaar.[28] Door bij interventiestudies zowel een ziektespecifiek als een generiek instrument toe te passen, kan men de voordelen van beide soorten instrumenten uitbuiten.

Praktische overwegingen

De wetenschappelijke eisen kunnen weleens op gespannen voet staan met de haalbaarheid van de meting van de ervaren gezondheidstoestand in de praktijk. Voor de praktische toepassing speelt de gebruiksvriendelijkheid van een vragenlijst een belangrijke rol: de lengte van de vragenlijst, de inzichtelijkheid en de moeilijkheidsgraad van de vragen, de manier waarop de lijst ingevuld moet worden, de scoring van de antwoorden en de stappen die nodig zijn voor de berekening van schaalscores.

Het belang van deze verschillende kenmerken is uiteraard voor een deel afhankelijk van de functie van de vragenlijst en van de concrete onderzoeksituatie. Zo zal men aan een vragenlijst die in het spreek-

uurcontact met de patiënt wordt gebruikt, andere eisen stellen dan aan een vragenlijst voor wetenschappelijk onderzoek. Een vragenlijst die in de huisartspraktijk gebruikt wordt, moet ingepast kunnen worden in de dagelijkse routine en zo min mogelijk beslag leggen op de tijd van de patiënt en van de arts of de assistente. Dat betekent dat de lijst snel en gemakkelijk in te vullen moet zijn. Wanneer de huisarts bovendien de vragenlijstinformatie onmiddellijk in het spreekuur wil benutten, moeten de schaalscores gemakkelijk rechtstreeks te verkrijgen zijn en eenvoudig geïnterpreteerd kunnen worden.

De moeilijkheidsgraad van vragenlijsten wordt bepaald door het aantal vragen, de formulering en de lengte van de vragen, de antwoordmogelijkheden en de eventuele routing (het beantwoorden van vervolgvragen afhankelijk van het antwoord op een voorgaande vraag). Het hangt van de moeilijkheidsgraad af of de vragenlijsten door een (goed getrainde) interviewer moeten worden afgenomen of door de patiënt zonder hulp kunnen worden ingevuld. De belasting die het beantwoorden van een vragenlijst veroorzaakt, is mede bepalend voor de bereidheid aan het onderzoek deel te nemen. Uit verschillende studies is overigens gebleken dat patiënten zonder bezwaren ook nogal omvangrijke lijsten, zoals de SIP invullen, wanneer ze de gevraagde informatie als zinvol en nuttig ervaren.[19] Het ervaren nut hangt onder meer af van de inzichtelijkheid van de vragen(lijst) en de relevantie van de vragen in de ogen van de patiënt. Ook de arts/onderzoeker zal zich rekenschap moeten geven van het nut (de 'clinical utility') van een meting van de ervaren gezondheid.

Boeken waarin vele vragenlijsten zijn beschreven

McDowell I, Newell C. Measuring health: a guide to rating scales and questionnaires. Oxford: Oxford University Press, 1987.

Walker SR, Rosser RM, eds. Quality of life: assessment and application. Lancaster: MTP Press, 1988.

Thompson C, ed. The instruments of psychiatric research. Chichester, etc.: John Wiley & Sons, 1989.

Wetzler S, ed. Measuring mental illness: psychometric assessment for clinicians. Washington DC: American Psychiatric Press, 1989.

Sartorius N, Goldberg D, De Girolamo G, et al., eds. Psychological disorders in general medical settings. Toronto: Hogrefe & Huber / WHO, 1990, i.h.b. Wittchen HU, Ahmoi Essau C. Assessment of symptoms and psychosocial disabilities in primary care (pp. 111-36).

Spilker B, ed. Quality of life assessment in clinical trials. New York: Raven Press, 1990.

Essink-Bot ML, Rutten-van Mölken MPMH. Het meten van de gezondheidstoestand. Rotterdam: Erasmus Universiteit Rotterdam, 1991.

Bowling A. Measuring health: a review of quality of life measurement scales. Buckingham: Open University Press, 1991.

Wilkin D, Hallam H, Doggett MA. Measures of need and outcome for primary health care. Oxford, etc.: Oxford University Press, 1992.

Evers A, Van Vliet-Mulder JC, Ter Laak J. Documentatie van tests en testresearch in Nederland. Assen, Maastricht: NIP/ Van Gorcum, 1992.

König-Zahn C, Furer JW, Tax B. Het meten van de gezondheidstoestand: beschrijving en evaluatie van vragenlijsten. I. Algemene gezondheid. Assen: Van Gorcum, 1993.

König-Zahn C, Furer JW, Tax B. Het meten van de gezondheidstoestand: beschrijving en evaluatie van vragenlijsten. II. Lichamelijke gezondheid, sociale gezondheid. Assen: Van Gorcum, 1994.

Furer JW, König-Zahn C, Tax B. Het meten van de gezondheidstoestand: beschrijving en evaluatie van vragenlijsten. III. Psychische gezondheid. Assen: Van Gorcum, ter perse.

Zoals vermeld worden bij de meeste vragenlijsten de antwoorden tot schaalscores samengevat. Het berekenen van deze schaalscores kan nogal bewerkelijk zijn. Dit is geen bezwaar, wanneer het geautomatiseerd kan gebeuren. In de patiëntenzorg kan echter een op zichzelf eenvoudige maar toch praktisch bewerkelijke scoreberekening onoverkomelijke problemen opleveren.

De meeste vragenlijsten zijn oorspronkelijk niet in Nederland ontwikkeld. Directe toepassing is daarom vaak niet mogelijk, omdat voor de vertaling strenge criteria moeten worden gehanteerd. Daarom zou eerst bekeken moeten worden of al een goede Nederlandse versie beschikbaar is en of deze uitvoerig is getest op validiteit en betrouwbaarheid. Bij een geautoriseerde vertaling mag men aannemen dat deze volgens de regels is uitgevoerd.

Op sommige vragenlijsten berust een copyright, waarin uiteenlopende eisen aan de potentiële gebruiker zijn vastgelegd. Deze vragenlijsten zullen (over het algemeen) op hun betrouwbaarheid en validiteit gecontroleerd zijn.

Aanbevelingen

Tegenwoordig kan de huisarts of onderzoeker uit een groot aantal vragenlijsten kiezen. Zoals uit deze bijdrage gebleken zal zijn, is dit aanbod zeer divers, en is er een grote kans dat men iets van zijn gading vindt. Dit rijke aanbod maakt de keuze echter niet eenvoudig. Om tot een weloverwogen beslissing te komen is het nuttig vooraf duidelijkheid te verkrijgen over de volgende vragen:

- Welke aspecten van de ervaren gezondheidstoestand zijn voor mijn onderzoeksvraagstelling relevant; wat moet beslist gemeten worden en wat zou verder nog interessant kunnen zijn?
- Bij welke doelgroep wil ik de ervaren gezondheidstoestand meten, bijvoorbeeld bij welke leeftijdsgroepen of bij welke patiëntencategorie?
- Welk type vragenlijst is het meest geschikt voor mijn doel: een generieke of een ziektespecifieke, of een combinatie van generiek en ziektespecifiek?

Zoek vervolgens naar een zo goed mogelijk geteste vragenlijst die past bij de vraagstelling en die geschikt is voor de beoogde onderzoekspopulatie. Informatie hierover is te vinden in boeken en artikelen. Van groot belang kan zijn dat de eigen resultaten vergeleken kunnen worden met nationale en internationale gegevens. Kies dan bij voorkeur een lijst die ook in andere landen gebruikt wordt.

Wat de beschikbare instrumenten waard zijn, is in de door ons aangehaalde literatuur te vinden. Meer en meer komen ook boeken ter beschikking waarin veel informatie over vragenlijsten is samengebracht. In het *kader* op pag. 115 wordt een overzicht gegeven van boeken waarin een groot aantal vragenlijsten is beschreven.

Literatuur

1 Spilker B, Molinek FR jr, Johnston KA, et al. Quality of life bibliography and indexes. Med Care 1990; 28 suppl to nr 12.
2 Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: development and final revision of a health status measure. Med Care 1981; 19: 787-805.
3 De Melker RA, Touw-Otten F, Jacobs HM, Luttink A. De waarde van de 'Sickness Impact Profile' als uitkomstmeting. Ned Tijdschr Geneeskd 1990; 134: 946-8.
4 Hunt SM, McEwen J, McKenna SP. Measuring health status. London: Croom Helm, 1986.
5 Erdman RA, Passchier J. The Dutch version of the Nottingham Health Profile: investigations of psychometric aspects. Psychol Reports 1993; 72: 1027-35.
6 Essink-Bot ML, Van Agt HME, Bonsel GJ. NHP of SIP: een vergelijkend onderzoek onder chronisch zieken. T Soc Geneesk 1992; 70: 152-9.
7 Meyboom-de Jong B, Smith RJA. Studies with the Dartmouth COOP Charts in general practice: comparison with the Nottingham Health Profile and the General Health Questionnaire. In: WONCA Classification Committee. Functional status measurement in primary care. New York, etc.: Springer, 1990.
8 Nelson EC, Landgraf JM, Hays RD, et al. The COOP Function Charts: a system to measure patient function in physicians' office. In: WONCA Classification Committee. Functional status measurement in primary care. New York, etc.: Springer, 1990.
9 Weel C van, Scholten JHG. De Dartmouth COOP Functional Health Assessment Charts/WONCA: een eenvoudig instrument om de functionele toestand van patiënten in de huisartspraktijk te meten. Huisarts Wet 1992; 35: 376-80.
10 Ware jr JE, Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36): I. Conceptual framework and item selection. Med Care 1992; 30: 473-83.
11 McHorney CE, Ware JE jr, Raczek AE. The MOS 36-item Short Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care 1993; 31: 247-63.
12 Zee K van der, Sanderman R, Heyink J. De psychometrische kwaliteiten van de MOS 36-item Short Form Health Survey (SF-36) in een Nederlandse populatie. T Soc Geneesk 1993; 71: 183-91.
13 Meenan RF, Gertman PM, Mason JH. Measuring health status in arthritis. The Arthritis Impact Measurement Scales. Arthritis Rheum 1980; 23: 146-52.
14 Meadows KA, Brown K, Thompson C, Wise PH. The Diabetes Health Profile (DHP): preliminary validation of a new instrument. Diabetic Med 1989; 6: suppl 2.
15 Guyatt GH, Berman LB, Townsend M, et al. A measure of quality of life for clinical trials in chronic lung disease. Thorax 1987; 42: 773-8.
16 Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. Appl Psych Meas 1977; 1: 385-401.
17 Swanborn PG. Schaaltechnieken. Meppel: Boom, 1982.
18 Dohrenwend BP, Dohrenwend BS. Perspectives on the past and future of psychiatric epidemiology. Am J Public Health 1982; 72: 2171-9.
19 Deyo RA, Patrick DL. Barriers to the use of health status measures in clinical investigation, patient care, and policy research. Med Care 1989; 27(3): S254-68.
20 American Psychological Association. Standards for educational and psychological tests. Washington: American Psychological Association, 1974.
21 Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol Bull 1959; 56: 81-105.
22 Hadorn DC, Hays RD. Multitrait-multimethod analysis of health-related quality-of-life measures. Med Care 1991; 29: 829-40.
23 Nunnally JC jr. Psychometric theory. New York: McGraw-Hill, 1978.
24 Jaeschke R, Guyatt GH. How to develop and validate a new quality of life instrument. In: Spilker B, ed. Quality of life assessments in clinical trials. New York: Raven Press, 1990.
25 Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chron Dis 1987; 40: 171-8.
26 Streiner DL, Norman GR. Health measurement scales. A practical guide to their development and use. Oxford: Oxford University Press, 1989.

Vervolg literatuur op pag. 128.

Abstract

König-Zahn C, Furer JW. Choosing a questionnaire – methodological and practical considerations. Huisarts Wet 1995; 38(3): 110-6, 128.

The assessment of perceived health status has met increasing interest in recent years. Nowadays a lot of questionnaires, meant to measure functional status, well-being or other aspects of perceived health, are available. Not all of these instruments do meet the required measurement standards. A deliberate choice of a questionnaire requires insight into the wide variety of available instruments as well as scrutiny of the scientific and practical requirements which should be met by an instrument. The choice of an appropriate questionnaire will mainly be guided by the particular research objective under study. If the objective is directed towards a specific condition or disease, a disease-specific questionnaire should be considered first. Nevertheless even in this situation a broad generic instrument can be very useful because it enables the investigator to compare different patient groups. A separate paragraph discusses the scientific requirements to be met by a questionnaire (i.e. validity, reliability, and sensitivity to change or responsiveness). A discussion of some practical issues and constraints is intended to round up the choice of the most appropriate instrument.

Key words Family practice; Health status assessment.

Correspondence Mrs. C. König-Zahn, Vakgroep Huisartsgeneeskunde, Sociale Geneeskunde en Verpleeghuisgeneeskunde, Katholieke Universiteit Nijmegen, Postbus 9101, 6500 HB Nijmegen, The Netherlands.