Triemstra, M., Hendriks, M., Delnoij, D., Rademakers, J. Doelspecifieke versies van CQ-index meetinstrumenten: korter, krachtiger, en specifieker meten? TSG: Tijdschrift voor Gezondheidswetenschappen: 2008, 86(8), 486-493

Doelspecifieke versies van CQ-index meetinstrumenten: korter, krachtiger, en specifieker meten?

MATTANJA TRIEMSTRA,[1] MICHELLE HENDRIKS,[1] DIANA DELNOIJ,[2] JANY RADEMAKERS [1]

ABSTRACT

CQ-index vragenlijsten meten de ervaringen van patie¨nten met de zorg. Over het algemeen zijn de lijsten lang om aan de uiteenlopende doelstellingen en informatiebehoeften van diverse partijen te voldoen. Dit brengt meer kosten voor dataverzameling en een risico op non-respons met zich mee. Ook sluiten de lijsten niet altijd goed aan op specifieke toepassingen in de praktijk. Voor een efficiënter gebruik van de CQ-index is het daarom de vraag of er verkorte modules of specifieke vragensets kunnen worden ontwikkeld die zijn toegesneden op de diverse gebruiksdoelen. Dit artikel biedt met statistische criteria een leidraad voor een nadere selectie van onderwerpen en specifieke vragen voor diverse doeleinden. Daarbij wordt grofweg onderscheid gemaakt in modules die gericht zijn op het in kaart brengen van verschillen tussen zorgaanbieders of instellingen, en versies die geschikt zijn voor het monitoren van verbetermogelijkheden in de zorg. Toepassing van de criteria wordt geïllustreerd aan de hand van de CQ-index Huisartsenzorg.

INLEIDING

Als standaardsystematiek voor het meten van klantervaringen in de zorg kent de CQ-index uiteenlopende gebruiksdoelen[1]. Zo moeten metingen informatie opleveren voor het opstellen van keuze-informatie voor consumenten, het ondersteunen van het zorginkoopbeleid van zorgverzekeraars, het signaleren van verbetermogelijkheden voor zorgaanbieders, en voor het leveren van (beleids) informatie aan patiënten-/cliëntenorganisaties, overheid en toezichthouders.

Om aan de uiteenlopende doelstellingen te voldoen, zijn CQ-index vragenlijsten over het algemeen lang en dit betekent meer kosten voor dataverzameling en een risico op non-respons. Ook is niet altijd duidelijk welke vragen voor welk doel geschikt zijn. Dit vraagt om de ontwikkeling van specifieke vragensets of modules.

LANGE VRAGENLIJSTEN VOOR MEERDERE DOELEINDEN

Niet zelden telt een CQI vragenlijst meer dan 100 vragen (zie Zuidgeest et al., 2008).[2] Dit betreft vooral sectorbrede of aandoeningspecifieke lijsten die over meerdere disciplines gaan. De betrokkenheid van verschillende partijen bij de ontwikkeling van een CQ-index vragenlijsten is medebepalend voor de lengte van lijsten. Patiënten-/cliëntenorganisaties, zorgverzekeraars en zorgaanbieders leveren gedurende een ontwikkeltraject input, als aanvulling op informatie uit focusgroepsgesprekken, standaarden/richtlijnen en literatuur.[3] Deze veelzijdige inbreng draagt bij aan lange conceptvragenlijsten die zelden korter uit een commentaarronde komen, zelfs niet als de diverse partijen nadrukkelijk worden verzocht om vragen te schrappen. De testfasen moeten dan uitsluitsel geven over welke vragen alsnog kunnen komen te vervallen. Wel worden aan het eind van een ontwikkeltraject vaak alsnog 10-30 vragen geschrapt die niet aan de maatstaven voor validiteit, betrouwbaarheid en vergelijkbaarheid voldoen. Lange vragenlijsten betekenen een grotere belasting en tijdsinvestering voor respondenten, wat ten koste kan gaan van de respons, representativiteit en validiteit van de antwoorden. Zeker bij oudere doelgroepen en mensen met lichamelijke of cognitieve beperkingen.[4] Kortere vragenlijsten gaan over het algemeen gepaard met een hogere respons.[5] Hoewel deze relatie vooralsnog niet is aangetoond voor de CQ-index,[2] zijn er voor verwante CAHPS®-lijsten wel aanwijzingen voor dit verband.[6] Lange vragenlijsten zijn ook duurder in toepassing en gebruik. Dit betreft hogere druk- en portokosten (zeker als bij een lage respons meer herinneringen me´t vragenlijsten moeten worden verstuurd) of hogere interviewkosten en meerkosten voor data-entry. Financiële consequenties zijn bij kleinschalige metingen nog te overzien, maar bij toepassing in grootschalige landelijke metingen lopen de kosten van dataverzameling hoog op. Dit alles roept om zo kort mogelijke vragenlijsten. Tevens roept de veelheid aan verzamelde informatie de vraag op wat nu precies de specifieke toepassingsmogelijkheden zijn. Dit vraagt om een verkenning van mogelijkheden voor verdere inkorting van vragenlijsten en de ontwikkeling van modules of vragensets die geschikt zijn voor specifieke doeleinden, of te wel doelspecifieke versies.

MODULES

Vooral bij grootschalige toepassing zoals in benchmarking, gericht op het vergelijken van prestaties van zorgaanbieders, is het aantrekkelijk om te werken met een selectie van onderwerpen en vragen. Zo'n 'vergelijkingsmodule' zou de basis kunnen vormen voor keuze-informatie voor consumenten, aangevuld met andere belangrijke kwaliteitsinformatie. De vergelijkingsinformatie en het inzicht in algemene verbeterpunten in een sector kunnen zorgverzekeraars gebruiken voor selectieve zorginkoop en de keuze van voorkeuraanbieders ('preferred providers'-contracten) . Voor het (interne) kwaliteitsbeleid van zorgaanbieders zijn juist vragensets geschikt die gedetailleerde management-/ stuurinformatie kunnen aanleveren over verbetermogelijkheden in de zorg, of die verbeterprogramma's kunnen evalueren. Met zo'n 'verbetermodule' en herhaalde metingen kunnen zorgaanbieders hun kwaliteitsverbeteringen monitoren. Vervolgens kunnen zij met zorgverzekeraars om tafel zitten om prestatie-afspraken te maken en 'pay-for-performance'-contracten af te sluiten . Met doelspecifieke versies kan mogelijk nog korter, krachtiger en specifieker worden gemeten. Overigens betekent de samenstelling van modules niet dat de vragensets ook echt afzonderlijk, in aparte metingen, dienen te worden uitgevraagd. De kracht van de CQ-index

schuilt immers vooral in het brede draagvlak dat ontstaat doordat het meerdere doelen dient en het instrument voor alle betrokken partijen interessante informatie oplevert .

DOEL EN OPZET VAN DIT ARTIKEL

Voor een efficiënte toepassing van de CQ-index is het dus wenselijk om kortere vragenlijsten en specifieke vragensets te ontwikkelen die meer zijn toegesneden op de gebruiksdoelen. Dit artikel beoogt daarom een methode te presenteren voor een nadere typering en selectie van vragen. Daarbij kunnen grofweg twee soorten vragensets of modules worden onderscheiden, gegeven de hiervoor genoemde toepassingen: a vragen die geschikt zijn voor het meten van verschillen tussen zorgaanbieders; b vragen die geschikt zijn voor het meten van verbeterpotentieel of kwaliteitsverbetering. Overlap tussen deze modules is waarschijnlijk. Het artikel presenteert (statistische) criteria voor de selectie van de zogeheten 'ervaringsvragen' (frequentie-, ja/nee- en probleemvragen). Eerst wordt de zoektocht naar selectiecriteria beschreven. De vraag 'Wat is een relevant of betekenisvol verschil?' vormde hierbij een rode draad. Immers, bij CQI metingen staat het aantonen van verschillen – tussen zorgaanbieders/verzekeraars of in de tijd – centraal. Vervolgens worden voor elke toepassing drie selectiecriteria gepresenteerd. De keuzes voor de criteria en bijbehorende afkappunten worden zoveel mogelijk met literatuur en empirische gegevens onderbouwd. Toepassing van de criteria wordt geïllustreerd aan de hand van de CQ-index Huisartsenzorg Overdag.[7]

Criteria voor de selectie van vragen

Voor een nadere selectie van vragen is gezocht naar relevante universele, robuuste, valide en praktisch toepasbare criteria die op een eenduidige wijze op verschillende CQI lijsten kunnen worden toegepast. Uitgangspunten hierbij waren: . criteria moeten over alle vragenlijsten, schalen en afzonderlijke vragen heen kunnen worden gelegd; . meest (aan)sprekende of voor de hand liggende maten (face validity); . meest gevoelige of onderscheidende maten (sensitiviteit); . praktische uitvoerbaarheid en gebruiksgemak bij dataanalyses. De keuze van criteria en bijbehorende afkappunten is zoveel mogelijk gebaseerd op literatuur en bevindingen uit vergelijkbaar onderzoek. Statistische criteria zijn grofweg onder te verdelen in geankerde (anchor-based) maatstaven en op verdelingen gebaseerde (distribution based) relatieve maatstaven.8 Een criterium is relatief als er geen ankerpunt - een ex- terne valide maat of norm - is waaraan het refereert (zoals een minimaal verschil in een relevante klinische of patiëntgerelateerde uitkomstindicator). Patiëntervaringen zoals gemeten met de CQ-index zijn vooralsnog niet geankerd omdat een externe validatie en normen over wat patiënten relevante verschillen vinden tot dusver ontbreken. Zo weten we bijvoorbeeld nog niet hoe groot verschilscores voor patiënten moeten zijn om te spreken over 'betere' en 'slechtere' zorgaanbieders. Daarom zijn de hier gepresenteerde criteria uitsluitend statistische maten, gebaseerd op de metrische eigenschappen van variabelen of op algemene vuistregels of referentiewaarden uit de literatuur. De zoektocht naar betekenisvolle criteria en afkappunten startte met een literatuuronderzoek naar relevante studies op het gebied van kwaliteit van zorg vanuit het patiënten-/cliëntenperspectief, patiëntgerelateerde uitkomsten (patient-reported outcome, kortweg PRO), en betekenisvolle verschillen (Minimal Important Difference, kortweg MID). Dit wees uit dat er op het gebied van patiëntervaringen met de zorg nog geen algemene criteria of vuistregels voor het vaststellen van betekenisvolle verschillen lijken te bestaan. Wel doen collega-onderzoekers in de Verenigde Staten in de zogenaamde CAHPS1-kit9 twee suggesties voor het definiëren van relevante verschillen in patiëntervaringen. De eerste methode betreft het benoemen van een procentueel verschil tussen het algemene gemiddelde en de dichtstbijzijnde grenswaarde van een schaal. De tweede methode betreft het benoemen

van een betekenisvol absoluut verschil tussen een score en het algemene gemiddelde. Beide methoden vereisen nog wel consensus voor het bepalen van afkappunten, en hier laat de CAHPS1-kit onderzoekers verder in het ongewisse. Onderzoek op het gebied van gezondheidsuitkomsten en kwaliteit van leven biedt wel nadere aanknopingspunten. Zo blijkt een klinisch relevant verschil of MID doorgaans te worden bepaald aan de hand van: 1) de effect size (ES), waarbij een MID kan variëren tussen 0.20-0.30 ES;10 2) de standaarddeviatie (SD), waarbij de MID overeenkomt met de respectievelijk door Cohen (1988), Sloan et al. (2005) en Norman et al. (2003) gesuggereerde 20%, 33% of 50% SD;11,12,13 3) absolute verschilscores, zoals de door Jaeschke et al. (1989) geopperde 0,5 punten op een 7-puntsschaal voor verbetering of verslechtering);14 of 4) de standaardmeetfout (zie Wyrwich, 1999).15 Na vergelijking van deze verschillende methoden voor het schatten van een MID stelden Ringash et al. (2007) als praktische vuistregel voor om 10% van de schaalrange als betekenisvol verschil te zien.16 Uiteindelijk zijn voor elke toepassing van de CQ-index (a. verschillen meten; b. kwaliteitsverbetering) drie criteria geformuleerd voor de selectie van vragen. Tabel 1 presenteert deze criteria met afkappunten. Hieronder volgt een nadere omschrijving en verantwoording van de criteria en bijbehorende afkappunten.

A VERSCHILLEN METEN

Zorgaanbieders kunnen onderling worden vergeleken op het niveau van praktijken, afdelingen, zorgnetwerken of zorgregio's. Deze analyse-eenheden worden verder kortweg aangeduid als AE. Voor deze module zijn de volgende drie criteria geselecteerd: 1 de intra-klasse correlatie (ICC) als maat voor het discriminerend vermogen; 2 de spreiding van AE-scores over drie klassen of 'sterrencategorieën'; 3 maximaal verschil tussen beste en slechtste AE-score. Omdat vergelijkende informatie voor case mix gecorrigeerd moet zijn, worden deze criteria bij voorbaat toegepast op de voor leeftijd, opleiding en gezondheid (en eventueel ook voor geslacht of andere variabelen) gecorrigeerde scores. De criteria en afkappunten worden hieronder toegelicht.

Intra-klasse correlatie (ICC)

Bij het vaststellen van verschillen gaat het primair om het discriminerend vermogen zoals dat ook in het ontwikkeltraject van een CQ-index wordt getest.3,17 Dit betreft het vermogen om onderscheid te maken tussen zorgaanbieders of zorgverzekeraars. CQI metingen laten vooralsnog zien dat de intra-klasse correlatie (ICC), ofwel de variatie tussen zorgaanbieders of organisatorische eenheden, per lijst en onderwerp aanzienlijk verschilt. Significante ICCs voor de gerapporteerde schalen variëren van 1,1% tot 38,4% en zijn gemiddeld 7,4%.18 De medi- aan voor significante ICCs bedraagt 4,9%. Dus gemiddeld gesproken wordt ruim 7% van alle variatie verklaard door verschillen tussen zorgaanbieders/instanties, en voor de helft van de schalen is deze variatie minstens 5%. De items of schalen van een lijst dienen in ieder geval te worden geselecteerd op basis van een significante ICCs, aangezien dan met zekerheid verschillen tussen eenheden kunnen worden aangetoond. Dit komt neer op ICCs vanaf 1%, aangezien de tot dusver gerapporteerde significante ICCs voor CQ-index vragenlijsten groter zijn dan 1,1%.18 Voor een strengere selectie kan worden gekozen voor een afkappunt op basis van de algemene spreiding van ICCs, bijvoorbeeld de kwartielscore of mediaan welke tot 25% of 50% minder vragen zullen leiden. Zo zal toepassing van de mediaan (dus significante ICCs vanaf 5%) resulteren in modules die doorgaans nog maar de helft van het oorspronkelijke aantal vragen bevatten. Maar omdat dit voor sommige vragenlijsten een nogal streng criterium zal blijken te zijn, is het aan te bevelen om van de ICC-kwartielwaarde uit te gaan welke tot nu toe neerkomt op 2,5% (afkappunt: significante ICCs >2,5%).

Spreiding van scores over drie categorieën

Het onderscheidend vermogen kan ook worden bekeken aan de hand van de verdeling van AE-scores over drie statistisch te onderscheiden groepen of 'sterrencategorieën ¨n':1 bovengemiddeld (***), gemiddeld (**), of benedengemiddeld (*). Dit is toegepast in eerder onderzoek naar mogelijkheden voor het inkorten van een enquête voor het ziekenhuisvergelijkingssysteem.19 Brouwer en Delnoij (2004) stelden dat een score onderscheidend was als minder dan 70% van de ziekenhuizen hier 'gemiddeld' op scoorde. Deze min of meer arbitraire keuze bleek bruikbaar voor verdere inkorting van de ziekenhuisvragenlijst. Door dit criterium toe te passen op CQI metingen kunnen met de geselecteerde vragen – uitgaande van een normale verdeling – de circa 15% significant best presterende (en 15% slechtste) zorgaanbieders worden geïdentificeerd. Dit is van belang voor het vaststellen van 'best practices' en voor het afsluiten van 'preferred provider contracten' door zorgverzekeraars.

Maximaal verschil tussen beste en slechtste score

Aanvullend op de voorgaande twee criteria, die iets zeggen over variatie en spreiding, zegt het absolute verschil tussen de 'allerbeste' AE en de 'allerslechtste' AE (Dmax = score 'best practice' - score 'worst practice') iets over de maximale grootte van het verschil. Als we de vuistregel van Ringash et al. (2007) als uitgangspunt nemen voor een betekenisvol verschil, dan geldt 10% van de schaalrange als ondergrens.16 Dit betekent dat voor de frequentievragen (4-punts antwoordschalen, 1-4) het verschil groter moet zijn dan 0,3, voor de probleemvragen (1-3) geldt meer dan 0,2, en voor de ja/nee-vragen meer dan 0,1. Bij waarderingscijfers (0-10) zouden volgens dit criterium alleen verschillen groter dan 1 relevant zijn.

B. KWALITEITSVERBETERING

Bij deze module gaat het om het meten van verbeterpotentieel en kwaliteitsverbetering. In CQ-index rapportages komt dit onderwerp aan de orde in de vorm van 'verbeterscores' en een top 10 van verbetermogelijkheden. Maar de grootte van verschillen tussen analyse-eenheden is hier ook van belang. Voor deze module zijn daarom de volgende criteria geselecteerd: 1 het gemiddelde verbeterpotentieel; 2 het percentage patiënten/cliënten met een suboptimale ervaring; 3 de verbeterscore. Omdat informatie over het verbeterpotentieel inzicht moet geven over optimale afstemming van het zorgaanbod op de 'lokale' vraag en deze informatie primair bedoeld is voor het kwaliteitsbeleid van zorgaanbieders, kunnen de criteria op ruwe data (ongecorrigeerde scores) worden toegepast. De drie criteria en mogelijke afkappunten worden hieronder verder toegelicht.

Gemiddelde verbeterpotentieel

Hier gaat het om het verschil tussen het algemene groepsgemiddelde en het gemiddelde van alle 'best practices' die bovengemiddeld scoorden: D_{mean} = gemiddelde 'best practices' (*** AEs) - groepsgemiddelde (alle AEs.). Dit sluit aan op de opvatting dat voor kwaliteitsverbetering realistische, haalbare 'best in class' prestatieniveaus ('benchmarks') moeten worden gedefinieerd.[20,21] Als bijbehorend afkappunt kan een verschil groter dan 5% van de schaalrange worden gekozen. Dit strookt met de conclusie van Ringash et al. (2007) dat voor het meten van 'verbeteringen' 5% van de schaalrange als ondergrens kan worden gehanteerd.[16] Voor de frequentievragen in de CQ-index vragenlijsten (1-4) betekent dit een gemiddeld verschil van meer dan 0,15, voor de probleemvragen (1- 3) meer dan 0,10, en voor de ja/nee-vragen >0,05. Voor waarderingscijfers (0-10) zou dit een verschil van tenminste 0,50 betekenen.

Percentage suboptimale ervaringen

Als meer dan 10% van de patiënten/cliënten 'nooit/soms', 'klein/groot probleem' of 'nee' antwoordt op ervaringsvragen (frequentie-, probleem- of ja/nee-vragen), moeten kwaliteitsverbeteringen worden overwogen. Dit min of meer arbitraire afkappunt geeft de mogelijkheid om zorg kritisch te evalueren.

Verbeterscore

Uit percentages suboptimale ervaringen en gemiddelde belangscores kunnen verbeterscores worden afgeleid.[3] Het product van de fractie suboptimale ervaringen (% negatieve ervaringen/100) en de bijbehorende belangscore (1-4) vormt de verbeterscore (0-4). Uitgaande van het voorgenoemde criterium van >10% suboptimale ervaringen en gemiddeld 'belangrijke' aspecten (belangscore 3) moet de verbeterscore dan minimaal 0,3 zijn.

DOORSLAGGEVENDE CRITERIA VOOR DE SELECTIE VAN VRAGEN

Bij de uiteindelijke selectie van vragen zijn vijf aanvullende voorwaarden doorslaggevend:

1 verplichte standaardvragen dienen te worden overgenomen (zie Handboek CQ-index; voor de samenstelling van vragenlijsten moet de selectie worden aangevuld met zeker 10 verplichte ervaringsvragen en waarderingscijfers, en 12 introductie-/achtergrondvragen);

2 als een schaal aan de criteria voldoet, dienen in ieder geval de twee hoogstladende items (met de sterkste item-totaal-correlaties) te worden geselecteerd;

3 items die volgens patiënten tot de top-10 van belangrijkste aspecten gerekend kunnen worden, moeten worden geselecteerd (als borging van het patiëntenperspectief);

4 onderwerpen of vragen die eerder zijn geselecteerd omdat ze aan de orde moeten komen bij keuze-informatie, maatschappelijke verantwoording of toezichtinformatie, worden naar mogelijkheid behouden (bijvoorbeeld zorginhoudelijke indicatoren uit richtlijnen, Etalage-plus informatie, of indicatoren uit normen voor verantwoorde zorg);

5 uiteindelijke keuzes over de selectie van onderwerpen en vragen dienen in overleg tussen inhoudsdeskundigen, eindgebruikers van de CQ-index, en onderzoekers tot stand te komen.

VOORBEELD VOOR TOEPASSING VAN DE CRITERIA

De criteria voor module A en B zijn ter illustratie toegepast op de in 2006-2007 ontwikkelde CQ-index Huisartsenzorg Overdag.[7] Deze lijst bestond aanvankelijk uit 81 items (62 ervaringsvragen) en na een pilot bij 32 huisartspraktijken in de regio's Drenthe en Rotterdam is een 'definitieve' lijst opgesteld met 66 vragen (52 ervaringsvragen). Hoewel het hier om een relatief korte en – in samenspraak met de vertegenwoordigers van de beroepsgroep, patiënten-/ consumentenorganisaties, verzekeraars en andere deskundigen – al aanzienlijk ingekorte CQ-index
vragenlijst gaat, is toch voor deze vragenlijst gekozen omdat de benodigde gegevens gepubliceerd zijn in het betreffende ontwikkelingsrapport.[7] Voor andere lijsten zijn de benodigde gegevens minder volledig gedocumenteerd. Tabel 2 geeft een overzicht van de schalen en items van de vragenlijst over huisartsenzorg die na toepassing van de criteria geschikt blijken te zijn voor doelspecifieke toepassingen. De geselecteerde schalen en vragen voldoen per doel aan minstens twee van de drie criteria. Bovendien is rekening gehouden met de eerste drie aanvullende voorwaarden (zie 'Doorslaggevende criteria'). Duidelijk wordt dat voor het meten van verschillen tussen huisartspraktijken 41 ervaringsvragen en voor het monitoren van kwaliteitsverbeteringen 42 ervaringsvragen geschikt zijn. Alle schalen over de praktijkorganisatie (Toegankelijkheid, Assistente) en de huisarts (Zorg op maat, Bejegening, Communicatie) blijken relevant voor het meten van verschillen. Voor

kwaliteitsverbetering zijn alleen de schalen Toegankelijkheid en Zorg op maat door de huisarts of andere zorgverleners (praktijkondersteuner of gespecialiseerd verpleegkundige) van belang. Daarnaast zijn per module een aantal losse vragen relevant; dit betreft vooral verplichte CQ-items (met name over bejegening) of volgens patiënten zeer belangrijke aspecten. Aan deze ervaringsvragen moeten vervolgens nog 14 standaardvragen (introductie-, selectie-, waarderings-, aanbevelings- en achtergrondvragen) worden toegevoegd, resulterend in modules met respectievelijk 55 en 56 vragen. Conclusie is dat toepassing van de criteria op de pilotversie van de huisartsenzorg-vragenlijst resulteert in twee even lange en sterk overlappende modules die samen een niet al te lange vragenlijst vormen. De definitieve lijst heeft een acceptabele lengte en invulduur (66 vragen in circa 15 minuten) en biedt het voordeel van e´e´n keer meten, twee keer weten. De uiteindelijke CQ-index Huisartsenzorg Overdag is een evenwichtige lijst die voor beide doelen tegelijk kan worden toegepast.

BESCHOUWING

De gepresenteerde criteriaset is primair bedoeld als handreiking voor onderzoekers die CQI meetinstrumenten ontwikkelen, zodat zij in staat worden gesteld om vragenlijsten eventueel verder in te korten en vragen te typeren en te selecteren voor doelspecifieke toepassingen. Hiertoe zal in het handboek van de CQ-index[17] een aanvullende werkinstructie worden opgenomen. Tevens kunnen de criteria ondersteuning bieden bij de afwegingen van inhoudsdeskundigen en diverse eindgebruikers over de relevantie en gebruiksdoelen van specifieke onderwerpen en vragen. Probleem bij het samenstellen van de criteria was wel het gebrek aan gegevens over voor patiënten betekenisvolle verschillen en normatieve referentiewaarden. Zolang criteria niet met ‘patientnormen’ geankerd zijn, blijven de afkappunten min of meer arbitrair. Des te belangrijker is het dat inhoudsdeskundigen, eindgebruikers en onderzoekers overeenstemming bereiken over de te selecteren vragen. Het toepassen van eenduidige criteria dient het algemene doel van de CQ-index als meetstandaard om klantervaringen op uniforme wijze in uiteenlopende zorgsituaties te meten, zodat de resultaten ook kunnen worden vergeleken. Daarom is voor de specifieke toepassingen ook gezocht naar universele, robuuste criteria. Probleem van deze algemene criteria is wel dat dit niet optimaal recht doet aan de verschillen tussen doelgroepen en settings. Een betekenisvol verschil kan immers per populatie en context verschillen.[22] Toch verdient het – gezien het ‘standaardisatiedoel‘ van de CQ-index – de voorkeur om de geformuleerde criteria uniform toe te passen. Toepassing van e´e´n en dezelfde afkappunten zal per lijst wel verschillend uitpakken. Zo zal het ICC-criterium (>2,5%) voor het meten van verschillen bij de vragenlijsten over Staaroperaties (met significante ICCs voor schalen varierend van 2 tot 3%),[23] Heup-/ knie-operaties (ICCs 2-4%)[24] en Diabetes (ICCs 1-4%)[25] maar weinig schalen en items selecteren. Het gepresenteerde voorbeeld voor de CQ-index Huisartsenzorg is in dit verband niet representatief, aangezien deze lijst veel verschillen in patientervaringen tussen huisartspraktijken laat zien. Terwijl significante ICCs voor de schalen in deze lijst gemiddeld 8,2% waren, was dit voor alle CQI meetinstrumenten gemiddeld 7,4%.[18] Toepassing van de criteria resulteerde bij de CQ-index Huisartsenzorg in twee even lange, sterk overlappende modules, maar bij andere vragenlijsten levert het waarschijnlijk een relatief korte ‘verschilmodule’ en een (iets) langere ‘verbetermodule’ op. Als toepassing van de criteria uitsluitend zou resulteren in kortere maar sterk overlappende modules, ontstaat wel de vraag wat nu precies de meerwaarde is boven andere methoden voor itemselectie. Als het doel puur en alleen de inkorting van vragenlijsten betreft, zijn andere methoden (zoals Item Respons Theorie, het Rasch model) wellicht meer geeigend. De voorgestelde criteria hebben juist als meerwaarde dat ze direct aansluiten op de analyses en rapportages van CQ-index metingen en toegespitst zijn op de gebruiksdoelen. Rest nog de vraag hoe de criteria per doel precies samenhangen. De exercitie voor de CQ-index

Huisartsenzorg wees uit dat er grote overeenstemming is tussen de drie criteria bij toepassing van de aanbevolen afkappunten. Daarbij pakten de criteria voor discriminerend vermogen (significante ICCs >2,5%) en het gemiddelde verbeterpotentieel (D_{mean} >0,15) relatief streng uit, omdat ze het minst aantal schalen en items selecteerden. De criteria D_{max} (>0,3) en het percentage suboptimale ervaringen (>10%) bleken daarentegen de meest sensitieve criteria. De eis om per doel aan minstens twee criteria te voldoen, bleek goed uit te pakken omdat bij module A het spreidingscriterium 'minder dan 70% gemiddeld' nauw aansloot op de D_{max}, en omdat bij module B het '% suboptimale ervaringen' en de verbeterscore vaak hetzelfde uitwezen. Deze bevindingen bevestigen de bruikbaarheid van de criteria en afkappunten. Tot slot. De voorgestelde criteriaset kan worden gebruikt voor verdere evaluatie van ontwikkelde CQ-index vragenlijsten, maar ook bij de ontwikkeling van nieuwe meetinstrumenten voor een nadere selectie van items - als aanvulling op methodologische argumenten omtrent validiteit, betrouwbaarheid en discriminerend vermogen. Het verdient vooral aanbeveling voor zeer lange lijsten, voor een (eventuele) verhoging van de respons en een vermindering van kosten voor dataverzameling.

LITERATUUR

1. Delnoij DMJ, Hendriks M, Gorp Kvan. De CQ-index: het meten van klantervaringen in de zorg. Tijdschr Gezondheidswet 2008; 86:440-46.
2. Zuidgeest M, Boer D de, Hendriks M, Rademakers J. Verschillende dataverzamelingsmethoden in CQI onderzoek: een overzicht van de respons en representativiteit van respondenten. Tijdschr Gezondheidswet 2008;86:455-62.
3. Rademakers J, Sixma H, Triemstra M et al. De constructie vanCQ-index meetinstrument: voorbeelden uit de praktijk. Tijdschr Gezondheidswet 2008;86:447-54.
4. Wijngaarden B van, Sixma H, Kok I. De bruikbaarheid van een CQ-index voor de GGZ en gehandicaptenzorg: specifieke aandachtspunten. Tijdschr Gezondheidswet 200886:463-70.
5. Edwards P, Roberts I, Clarke M et al. Methods to increase response rates to postal questionnaires. Cochrane Database Syst Rev 2007;2:MR000008
6. Gallagher PM, Fowler FJ. Size Doesn't Matter. Response Rates of Medicaid Enrollees to Questionnaires of Various Lengths. Presentation delivered at the 4th National CAHPS User Group Meeting, Baltimore, MD, 1998.
7. Meuwissen LE, Bakker DH de. CQ-index huisartsenzorg: kwaliteit vanuit het perspectief van patie¨nten. Meetinstrumentontwikkeling. Utrecht: NIVEL, 2008.
8. Broz . ek JL, Guyatt GH, Schu¨nemann HJ. How a well-grounded minimal importance difference can enhance transparency of labelling claims and improve interpretation of a patient reported outcome measure. Health Qual Life Outcomes 2006; 4:69.
9. CAHPS 2.0 Survey and Reporting Kit. CAHPS 3.2 Analysis Appendix, p. 9-10. (www.cahps.ahrq.gov)
10. Revicki DA, Cella D, Hays RD, Sloan JA, Lenderking WR, Aaronson NK. Responsiveness and minimal important differences for patient reported outcomes. Health Qual Life Outcomes 2006;4:70.
11. Cohen J. Statistical Power Analysis for the Behavioral Sciences (2nd edition). Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.
12. Sloan JA, Cella D, Hays RD. Clinical significance of patientreported questionnaire data: Another step toward consensus. J Clin Epidemiol 2005;58:1217-9. 13. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care 2003;41:582-92.
14. Jaeschke R, Singer J, Guyatt GH. Measurement of health status: Ascertaining the minimal clinically important difference. Control Clin Trials 1989;10:407-15.
15. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. J Clin Epidemiol 1999;52:861-73.

16. Ringash J, O'Sullivan B, Bezjak A, Redelmeier DA. Interpreting clinically significant changes in patient-reported outcomes. Cancer 2007;110:196-202.
17. Sixma H, Delnoij D (red.). Handboek CQI meetinstrumenten: een handleiding voor de ontwikkeling en het gebruik van Consumer Quality Index (CQI) vragenlijsten. Utrecht: Centrum Klantervaring Zorg, 2007.
18. Hendriks M, de Boer D, Spreeuwenberg P, Rademakers J. De CQindex: het vergelijken van klantervaringen binnen en tussen zorgsectoren. Achtergronddocument bij de Jaarrapportage Klantervaringen in de Zorg 2007 van het Centrum Klantervaring Zorg. Utrecht: CKZ/NIVEL, 2008.
19. Brouwer W, Delnoij DMJ. Aanpassing patie¨ntenenqueˆte ziekenhuisvergelijkingssysteem. Utrecht: NIVEL, 2004.
20. Kiefe CI, Weismann NW, Allison JJ, Farmer R, Weaver M, Williams OD. Identifying achievable benchmarks of care: concepts and methodology. Int J Qual Health Care 1998;10:443-7.
21. ABC^2 Handleiding 2005. (http://main.uab.edu/show.asp?- durki=14527)
22. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol 2008; 61:102-109.
23. Stubbe J, Dijk L van. Het discriminerend vermogen van de CQindex Staaroperatie. Utrecht: NIVEL, 2007.
24. Stubbe J, Dijk L van. Het discriminerend vermogen van de CQ index Heup-/Knieoperatie. Utrecht: NIVEL, 2007. 25. Stubbe J, Spreeuwenberg P, Asbroek G ten. CQ-index Diabetes: schaalconstructie, betrouwbaarheid en discriminerend vermogen van de ervaringenvragenlijst. Utrecht: NIVEL, 2007.