COSMIN: Development and evaluation of a checklist to assess the methodological quality of studies on measurement properties

Mokkink, L.B.

2010

document version
Publisher's PDF, also known as Version of record

Samenvatting

COSMIN

COSMIN is een acroniem voor 'COnsensus-based Standards for the selection of health Measurement INstruments'. Het COSMIN initiatief heeft als doel om de richtlijnen voor het selecteren van meetinstrumenten die zich op gezondheid richten te verbeteren.
Het onderzoek dat in dit proefschrift is beschreven draagt bij aan het doel van het COSMIN initiatief. We hebben ons in dit onderzoek gericht op wat we *evaluatieve* gezondheids-gerelateerde patiënt-gerapporteerde uitkomstmaten noemen, in het Engels heet dat health-related patient-reported outcomes (HR-PROs). Een instrument is evaluatief wanneer het wordt toegepast in een longitudinale studie om veranderingen over de tijd te meten. Een PRO is een uitkomstmaat die alleen door de patiënt zelf gerapporteerd kan worden, bijvoorbeeld over de kwaliteit van leven, pijn, functioneren of moeheid. Veelal worden deze constructen met een vragenlijst, interview of dagboekje gemeten.

Gezondheidsgerelateerde meetinstrumenten worden veelvuldig in wetenschappelijk onderzoek en in de klinische praktijk gebruikt. Bovendien zijn er vele vergelijkbare meetinstrumenten beschikbaar. Hieruit zal een keuze gemaakt moeten worden. De selectie van een meetinstrument hangt (idealiter) af van onder andere de kwaliteit van de meetinstrumenten waaruit gekozen kan worden. Om de kwaliteit van een meetinstrument te bepalen moeten de meeteigenschappen van het instrument onderzocht worden. Er zijn een aantal verschillende meeteigenschappen, zoals onder andere de betrouwbaarheid (reliability), de meetfout (measurement error), en de inhoudsvaliditeit (content validity). Studies waarin deze meeteigenschappen worden onderzocht moeten van methodologisch hoge kwaliteit zijn, zodat er een minimale kans is op vertekening (bias) van de resultaten. Resultaten die verkregen worden in studies die methodologisch zwak zijn, zijn moeilijk te vertrouwen. Om potentiële bias op te sporen, is het nodig de methodologische kwaliteit van studies naar meeteigenschappen te beoordelen.

Dit proefschrift gaat over de ontwikkeling en de evaluatie van de COSMIN checklist, die gebruikt kan worden om de methodologische kwaliteit te beoordelen van studies die een of meer meeteigenschappen onderzoeken. De COSMIN checklist bevat standaarden over de methodologische kwaliteit die ontwikkeld zijn in de COSMIN Delphi studie.
De onderzoeksvragen van deze studie waren:

1. Welke meeteigenschappen moeten onderzocht worden bij het beoordelen van evaluatieve HR-PROs, en hoe moeten die gedefinieerd worden?
2. Hoe moeten deze meeteigenschappen onderzocht worden? Met andere woorden, waar moet het design van de studie aan voldoen, en wat zijn goede statistische methoden om de meeteigenschappen te analyseren? Dit noemen wij de standaarden.
3. Welke criteria moeten worden toegepast om te bepalen of een meeteigenschap dan goed is?

Naast deze Delphi studie is in dit proefschirft ook een systematische review naar klinimetrische reviews, en een interbeoordelaarsbetrouwbaarheidsstudie naar de kwaliteit van de COSMIN checklist beschreven.

In hoofdstuk 1 wordt een inleiding over richtlijnen voor kwaliteitsbeoordelingen gegeven. Tevens wordt de opbouw van het proefschrift inclusief de onderzoeksvragen van de COSMIN Delphi studie beschreven.
In Hoofdstuk 2 presenteren we het design en onderzoeksvragen van de COSMIN Delphi studie. Het design van de Delphi studie bestond uit een voorbereidingsfase, vier schriftelijke Delphi rondes om tot consensus te komen over de onderzoeksvragen, en een evaluatie fase. De voorbereidingsfase van de Delphi studie bestond uit een systematische literatuur review naar klinimetrische reviews. Dit zijn reviews waarin de meeteigenschappen van evaluatieve gezondheidsmeetinstrumenten worden beschreven en beoordeeld. De vier schriftelijke rondes bestonden uit een vragenlijst en een feedback rapport van de vorige ronde. In dit hoofdstuk wordt beschreven hoe de experts voor de Delphi studie zijn geselecteerd. Deelnemende panelleden hadden een achtergrond in epidemiologie, statistiek, psychologie of geneeskunde. De panelleden werden gevraagd aan te geven in welke mate ze het met een voorstel eens waren. Dit konden ze aangeven op een 5-puntsschaal (sterk mee oneens, mee oneens, neutraal, mee eens, sterk mee eens). Wanneer 67% van de panelleden eens was met een voorstel beschouwden we dit als consensus. Tot slot werd in de evaluatie fase een interbeoordelaarsbetrouwbaarheidsstudie uitgevoerd.

In Hoofdstuk 3 beschrijven we de systematische review naar klinimetrische reviews. Doel van deze review was het *beoordelen* van de kwaliteit van het review proces van de klinimetrische reviews en het *beschrijven* van welke meeteigenschappen werden onderzocht in de klinimetrische reviews en hoe dat werd gedaan. Hiervan werd een inventarisatie gemaakt van welke meeteigenschappen werden beschreven en hoe deze werden gedefinieerd in de reviews, en welke standaarden werden gebruikt om de methodologische kwaliteit van studies te beoordelen en welke criteria werden gebruikt om de kwaliteit van de meetinstrumenten te beoordelen. Deze inventarisatie werd gebruikt als input in de COSMIN Delphi studie. De klinimetrische reviews werden gezocht in PubMed, Embase, en PsycInfo. We includeerden reviews die gingen over gezondheidsmaten die in een evaluatieve toepassing werden gebruikt en die als doel hadden om over de meeteigenschappen van deze reviews te rapporteren. Twee reviewers selecteerden de artikelen en extraheerden de resultaten onafhankelijk van elkaar.
We vonden 148 klinimetrische reviews over meetinstrumenten die algemene gezondheidsperceptie (43%), functionele status (21%), symptomen (17%), biologische en fysiologische processen (5%), of een combinatie van bovengenoemde concepten (14%) meten. In dit

hoofdstuk concluderen we dat het aantal klinimetrische reviews de laatste paar jaar substantieel is toegenomen. Echter, de kwaliteit van deze reviews laat nog veel te wensen over. In 56% (n=83/148) van de reviews werd de methodologische kwaliteit van de primaire studies (gedeeltelijk) beoordeeld door de auteurs en/of werden (sommige) meeteigenschappen geevalueerd. Met andere woorden, in 44% van deze reviews werden geen standaarden (voor het design van de studie en de statistische analyses) of criteria (voor de resultaten van de studie) toegepast voor een of meerdere meeteigenschappen. In deze 56% van de reviews die wel standaarden en/of criteria toepasten werd vaak maar een beperkt aantal standaarden of criteria toegepast. In slechts zeven reviews (5%) werden zowel standaarden als criteria voor alle meeteigenschappen toegepast.

De hoofdstukken 4 tot 6 zijn gebaseerd op de COSMIN Delphi studie. Deze internationale Delphi studie had als doel om tot consensus te komen (1) over welke meeteigenschappen relevant zijn bij het selecteren van evaluatieve HR-PROs, (2) over de terminologie en definities van deze meeteigenschappen, (3) over een taxonomie van meeteigenschappen, en (4) over standaarden voor het evalueren van meeteigenschappen. Terminologie, definities en standaarden die gebruikt waren in de klinimetrische reviews die wij in de systematisch review hadden geïncludeerd (hoofdstuk 3), of in een aanvullende literatuur search naar methodologische literatuur, werden gebruikt als input in de COSMIN Delphi studie.
In Appendix 4 is de COSMIN taxonomie gepresenteerd met daarin drie domeinen – betrouwbaarheid (reliability), validiteit (validity) en responsiviteit (responsiveness) – negen (aspecten van) meeteigenschappen en daarnaast ook interpreteerbaarheid (interpretability). In Appendix 5 kunnen de terminologie en definities gevonden worden. Een stroomdiagram met de vier-stappen procedure om de COSMIN checklist in te vullen is weergegeven in Appendix 6, en de COSMIN checklist wordt in Appendix 7 gepresenteerd.
We verwachten dat deze consensus zal leiden tot een uniformer gebruik van terminologie en definities in toekomstige literatuur over meeteigenschappen. Gebrek aan consensus heeft geleid tot verwarring over welke meeteigenschappen relevant zijn, welke concepten gepresenteerd zijn en hoe de meeteigenschappen het best geëvalueerd kunnen worden in termen van studie design en statistische methodes.

In Hoofdstuk 4 beschrijven we de consensus die het panel bereikt heeft over de terminologie (percentage consensus varieert tussen 74% en 88%, behalve voor één term: slechts 56% was het eens met de term structurele validiteit (structural validity)), de definities (percentage consensus varieert tussen de 68% en 88%) en een taxonomie van meeteigenschappen. Over sommige onderwerpen vond een uitgebreide discussie plaats. We beschrijven deze discussies van het panel over de positie van de meeteigenschappen interne consistentie (internal consistency) en responsiviteit (responsiveness) in de taxonomie, over de termen

'reliability' (betrouwbaarheid) en 'structural validity' (structurele validiteit) en tot slot over de definities voor de meeteigenschappen interne consistentie, betrouwbaarheid (reliability) en responsiviteit.

De discussie ging over of interne consistentie al dan niet beschouwd moest worden als een aparte meeteigenschap, of dat het een aspect was van de meeteigenschap betrouwbaarheid (reliability). Er werd besloten dat interne consistentie een aparte meeteigenschap was binnen het domein betrouwbaarheid (reliability). Daarnaast werd besloten dat het domein responsiviteit naast het domein validiteit zou bestaan, en we hebben bediscussieerd of het domein responsiviteit twee meeteigenschappen zou moeten omvatten, namelijk construct responsiviteit en criterion responsiviteit. Dit is vergelijkbaar met construct validiteit en criterion validiteit. Het panel wilde geen nieuwe termen introduceren, daarom heet de meeteigenschap responsiviteit. Het verschil tussen de standaarden voor situaties waarin een gouden standaard beschikbaar is en voor situaties waarin die niet beschikbaar is, komt nu tot uiting in de COSMIN standaarden voor responsiviteit.

Ook is discussie geweest over de keuze tussen de termen "reliability" (betrouwbaarheid) en "reproducibility" (reproduceerbaarheid) voor de meeteigenschap. Hoewel het domein de term "reliability" (betrouwbaarheid) heeft gekregen, had het panel ook de voorkeur voor het gebruik van de term "reliability" (betrouwbaarheid) voor de meeteigenschap. Daarnaast hebben we de keuze besproken tussen de termen "factorial validity" (factoriële validiteit) of "structural validity" (structurele validiteit). Argumenten waren dat "factorial validity" verwijst naar één van de methoden om deze meeteigenschap te evalueren, terwijl "structural validity" refereert naar het doel van dit aspect van de meeteigenschap construct validiteit. Het panel bereikte geen consensus over deze term. Daarom heeft de COSMIN Stuurgroep gekozen voor de term "structural validity", oftewel "structurele validiteit".

Om consensus te bereiken over een definitie voor interne consistentie, zijn we begonnen met een definitie waarin geprobeerd werd om het concept van interne consistentie uit te leggen en tegelijkertijd het verschil te benadrukken tussen interne consistentie en homogeniteit (ook wel unidimensionaliteit genoemd). Het panel had echter de voorkeur voor een definitie waarin alleen het concept van interne consistentie werd uitgelegd. Het verschil met homogeniteit is nu weergegeven in de standaarden voor interne consistentie.

De eerste definitie voor de meeteigenschap betrouwbaarheid (reliability) waarover we consensus bereikten, bleek niet in overeenstemming te zijn met de goede statistische methoden - intraclass correlation coefficient (ICC) en Cohen's kappa. Daarom werd een andere definitie voorgesteld en geaccepteerd door het panel (zie Appendix 5).

Het panel bereikte ook consensus over de definitie van responsiviteit (responsiveness): "the ability of an instrument to detect important change over time in the construct to be measured" (de mogelijkheid van een meetinstrument om belangrijke verandering over de tijd te detecteren in het construct dat wordt gemeten). Vervolgens werd besloten om het woord

'belangrijk' te verwijderen, omdat dit gaat over de interpretatie van de veranderscore, wat een ander issue is. Daarnaast is het afkappunt tussen wat wel en net niet een belangrijke verandering is arbitrair.

In Hoofdstuk 5 presenteren we de COSMIN checklist, waarin de standaarden zijn opgenomen waarover consensus is bereikt in de COSMIN Delphi studie voor het evalueren van de methodologische kwaliteit van studies naar meeteigenschappen. Er is een 4-stappen plan gemaakt om de COSMIN checklist in te vullen: in stap 1 wordt bepaald welke meeteigenschappen worden geëvalueerd in een artikel; in stap 2 wordt bekeken of er gebruikt wordt gemaakt van Item Response Theory (IRT) technieken; in stap 3 wordt de methodologische kwaliteit per onderzochte meeteigenschap geëvalueerd; stap 4 wordt gebruikt om te bepalen naar welke onderliggende populatie de resultaten van een studie gegeneraliseerd kunnen worden.
De COSMIN checklist heeft 12 boxen. Er is één box met vier items met algemene eisen voor artikelen die IRT gebruiken – te gebruiken bij stap 2. Er zijn 10 boxen met standaarden per meeteigenschap, waarbij bepaald kan worden of een studie voldoet aan de standaard voor goede methodologische kwaliteit – stap 3. Van deze 10 boxen zijn er negen voor de meeteigenschappen (interne consistentie, betrouwbaarheid (reliability), meetfout (measurement error), inhoudsvaliditeit (content validity), structurele validiteit (structural validity), hypothese toetsing (hypotheses testing), cross-culturele validiteit (cross-cultural validity), criterion validiteit (criterion validity), en responsiviteit (responsiveness); varierend van 5-18 items), en één box bevat de standaarden voor studies naar interpreteerbaarheid (interpretability; 9 items). De laatste box met algemene eisen voor de generalisatie van de resultaten van een studie wordt gebruikt bij stap 4. Voor een uitgebreide beschrijving over het gebruik van de checklist verwijzen we naar de COSMIN checklist handleiding op www.cosmin.nl.
Zoals eerder gezegd bevat de COSMIN checklist standaarden om de methodologische kwaliteit van studies naar meeteigenschappen te evalueren. Het is dus niet bruikbaar om de kwaliteit van een HR-PRO meetinstrument te evalueren. Daarvoor zijn criteria nodig om te bepalen wanneer de resultaten van studies naar meeteigenschappen goed genoeg zijn. Om uiteindelijk een meetinstrument te selecteren zijn deze criteria nodig en moeten in de toekomst nog verder ontwikkeld worden.

In Hoofdstuk 6 gaan we dieper in op een aantal items van de COSMIN checklist. We lichten onze keuzes voor geïncludeerde standaarden over design eisen en statistische methoden toe, tegen de achtergrond van bestaande literatuur. Doel hiervan is om een beter begrip van de redenen achter de items te bewerkstelligen, en daarmee de acceptatie en het gebruik van de COSMIN checklist te bevorderen.

We leggen uit dat interne consistentie alleen relevant is voor meetinstrumenten die constructen meten die gebaseerd zijn op reflectieve modellen, en dat homogeniteit (unidimensionaliteit) een voorwaarde is voor interne consistentie. Ook leggen we uit dat content validiteit gaat over relevantie en volledigheid. Het gaat hierbij om de manier waarop in een artikel een oordeel is gegeven over de relevantie van elk item en of de items samen het te meten construct volledig meten. Hypothese toetsing is een van de drie aspecten van de meeteigenschap construct validiteit. Het gaat over de relatie tussen scores op het meetinstrument en scores op andere meetinstrumenten of over verschillen in scores tussen relevante groepen. Het is een onophoudelijk, iteratief proces, waarbij specifieke hypotheses opgesteld en getoetst moeten worden over de te verwachte richting en grootte van de correlaties tussen (sub)schalen van meetinstrumenten of tussen verschillen in groepen. Criterion validiteit wordt gedefinieerd als 'the degree to which the scores of a HR-PRO instrument are an adequate reflection of a "gold standard"' (de mate waarin scores op een HR-PRO instrument een adequate weerspiegeling zijn van een "gouden standard"). Het Delphi panel bereikte consensus dat gouden standaarden voor HR-PROs niet bestaan, met als enige uitzondering dat wanneer een bestaande HR-PRO wordt verkort, de langere versie beschouwd kan worden als een gouden standaard voor de verkorte HR-PRO.
Responsiviteit wordt beschouwd als een aparte meeteigenschap, maar het panel kwam overeen dat het enige verschil tussen cross-sectionele (construct and criterion) validiteit en responsiviteit is dat validiteit refereert aan een single score en responsiviteit aan de verschilscore. Omdat er geen gouden standaarden bestaan voor HR-PROs, is de goede manier om responsiviteit te meten door a priori gedefinieerde hypotheses te toetsen over de relatie tussen veranderingen in scores op het meetinstrument en veranderingen in andere meetinstrumenten. In Hoofdstuk 6 hebben we tot slot uitgelegd waarom parameters zoals effect sizes (gemiddelde verschilscore / standaard deviatie (SD) op baseline), en aanverwante maten, zoals de standardised response mean (gemiddelde verschilscore / SD verschilscore) en ook maten als de gepaarde t-toets, Guyatt's responsiveness ratio (MIC/SD verschilscore bij stabiele patiënten) en minimal important change (MIC) ongeschikte maten zijn voor responsiviteit.

Het doel van de studie die in Hoofdstuk 7 is beschreven was om de interbeoordelaarsbetrouwbaarheid van de items uit de COSMIN checklist te onderzoeken. We hebben 75 artikelen waarin een of meerdere meeteigenschappen van HR-PROs werden onderzocht random geselecteerd. Van elk artikel hebben we de werklast voor de beoordelaar bepaald, gebaseerd op het aantal meeteigenschappen dat werd onderzocht, het aantal meetinstrumenten dat werd onderzocht, het aantal pagina's van het artikel en of er IRT werd toegepast. Achten-tachtig mensen hebben de methodologische kwaliteit van drie artikelen bepaald met behulp van de COSMIN checklist. Daarbij ontving elke deelnemer één artikel met een

relatief lage werklast, één met een gemiddelde werklast en één met een hoge werklast. Kappa's (one-way design) en het percentage overeenstemming werd berekend per item van de COSMIN checklist. In het algemeen waren de kappa's laag (59% van de items had een kappa onder de 0.40), terwijl de percentages overeenstemming acceptabel waren (dwz boven de 80% overeenstemming) bij 64% van de items. Oorzaken voor deze lage kappa's waren (1) een gebrek aan spreiding van de scores op de items; (2) de noodzaak bij een aantal items om een subjectief oordeel te geven; (3) gebrekkige rapportage van de artikelen waardoor er een tekort was aan informatie die nodig is om items te beantwoorden; (4) onzorgvuldig lezen van de COSMIN handleiding, of het oneens zijn met bepaalde items uit de COSMIN checklist; (5) gebruik van afwijkende terminologie en taxonomie door de auteurs van de te beoordelen artikelen; en (6) het moeilijke onderscheid tussen standaarden voor studies (die betrekking hebben op de methodologische kwaliteit van de studie) en de criteria voor resultaten (die betrekking hebben op de kwaliteit van de meetinstrumenten). Dit is met name lastig bij content validiteit.
Om de overeenstemming te verbeteren, hebben we de handleiding verbeterd. Ook raden we aan wanneer de COSMIN checklist gebruikt gaat worden (bijvoorbeeld in een klinimetrische review) om eerst wat ervaring te krijgen met het invullen van de COSMIN checklist. We raden sterk aan om de COSMIN taxonomie en terminologie te gebruiken. Bijvoorbeeld, als een auteur van een te beoordelen artikel een PRO instrument vergelijkt met een veel gebruikte andere PRO, zoals de SF-36, en dit beschouwt als criterion validiteit, dan raden wij aan om bij het evalueren van deze studie dit toch te beschouwen als een vorm van hypothese toesting (construct validiteit) en de bijbehorende box F in te vullen. Wanneer de checklist in een klinimetrische review wordt gebruikt, raden wij aan dat twee beoordelaars eerst onafhankelijk de checklist invullen, en daarna consensus bereiken over één eindoordeel. We adviseren het review-team om vooraf afspraken te maken over hoe de items die een subjectief oordeel nodig hebben gescoord zullen worden en hoe er omgegaan zal worden met gebrekkig gerapporteerde artikelen. De interbeoordelaarsbetrouwbaarheid van deze consensus verkregen door paren reviewers zou mogelijk hoger kunnen zijn, maar dit zou onderzocht moeten worden.

Naast het onderzoeken van de betrouwbaarheid van de COSMIN checklist, raden wij aan om de content validiteit, construct validiteit door middel van hypothese toetsing en de interpreteerbaarheid van de COSMIN checklist te onderzoeken. Interne consistentie, structurele validiteit, meetfout, cross-culturele validiteit, criterion validiteit en responsiviteit zijn volgens ons niet relevant voor de COSMIN checklist.

Hoofdstuk 8 bevat een algemene discussie van dit proefschrift. In dit hoofdstuk concluderen we dat de Delphi techniek een geschikte methode is om consensus te bereiken over

de COSMIN checklist en taxonomie, omdat deze methode met name geschikt is wanneer er een tekort is aan kennis of overeenstemming over een onderwerp. In een Delphi procedure kunnen voor- en tegenargumenten voor keuzes geïdentificeerd en overwogen worden. Daarnaast zijn beslissingen over terminologie en definities afhankelijk van de overeenstemming tussen experts, en deze beslissingen kunnen niet empirisch onderzocht worden. De plaats van elke meeteigenschap in de taxonomie is een logische consequentie van de gekozen definitie.

In dit hoofdstuk concluderen we vervolgens dat het panelleden die betrokken waren bij de Delphi studie een geschikt panel vormden, omdat vertegenwoordigers van alle disciplines die zich bezighouden met HR-PRO instrumenten hebben deelgenomen. Een groot aantal mensen dat bereid was om mee te werken aan de Delphi studie, bleven ook betrokken tot het einde van het Delphi proces, ondanks de grote tijdsbelasting die het vroeg.

De Delphi techniek is gevoelig voor drie soorten vertekening, namelijk selectie bias, subject bias en bias geïntroduceerd door de onderzoekers bij het interpreteren van de resultaten. Om selectie bias in de Delphi studie te vermijden, hebben we een heterogene groep panelleden uitgenodigd. Om subject bias te voorkomen, dat wil zeggen vertekening die ontstaat doordat panelleden met een gevestigde naam de meningen van de andere panelleden bewust of onbewust beïnvloeden, hebben we alle panelleden anoniem gehouden. Om interpretatie bias te voorkomen, hebben we van te voren het criterium voor consensus geformuleerd, en hebben we geprobeerd om zo transparant mogelijk te werken. Dit werd bereikt door de meningen van de panelleden te vragen, hen aan te sporen om de argumenten voor hun keuzes te geven, door alle antwoorden op vragen uit de vorige Delphi ronde in een feedback rapport terug te koppelen, en door onze keuzes uit te leggen.

Tot slot hebben we in dit hoofdstuk aanbevelingen gedaan voor vervolg onderzoek. Een van die onderwerpen was het combineren van scores per box, zodat een totaal score over de methodologische kwaliteit van een studie naar een meeteigenschap gegeven kan worden. Het COSMIN initiative, in samenwerking met de Klinimetrie Werkgroep van het EMGO institute for Health and Care Research, is bezig met het ontwikkelen van een scoringsysteem om de methodologische kwaliteit van studies te classificeren als uitstekend, goed, redelijk en slecht. Dit scoringssysteem moet nog worden geëvalueerd.

Om naast de methodologische kwaliteit van een studie te concluderen dat de kwaliteit van een meetinstrument voldoende is, moeten de studies waarin de meeteigenschappen zijn onderzocht ook goede resultaten hebben gevonden. Criteria om te bepalen wat *voldoende* of *goed* is zullen daarom – liefst gebaseerd op consensus – ontwikkeld moeten worden, en zo mogelijk ondersteund door empirische bevindingen.

Toelichting voor het gebruik van de COSMIN checklist kunnen worden gevonden op onze website, www.cosmin.nl.