Tool Analyses

Voor het vergelijken van patiëntervaringen
tussen zorgaanbieders | versie 1.0

Zorginstituut Nederland

nivel

Over deze tool

Deze tool bundelt kennis over hoe data van patiëntervaringen geanalyseerd kan worden zodat het gebruikt kan worden om zorgaanbieders te vergelijken. Deze vergelijkingsinformatie kan door zorgaanbieders en brancheorganisaties gebruikt worden als verantwoording aan patiënten, overheid en zorgverzekeraars.

De beoogde lezers van deze tool zijn statistici, methodologen en onderzoekers die zelf vergelijkende analyses uitvoeren. Deze tool biedt ook achtergrondinformatie aan zorgaanbieders (brancheorganisaties), patiëntenorganisaties en zorgverzekeraars die ondersteuning willen inhuren voor het uitvoeren van analyses voor het vergelijken van zorgaanbieders.

Deze tool maakt deel uit van een reeks van tools over patiëntervaringsmetingen, over onderwerpen als privacy, dataverzameling, en kwalitatieve methoden en evalueren en optimaliseren. De tools bieden verdiepende en praktische informatie in aanvulling op de Handreiking Ontwikkelen van Patiëntervaringsvragenlijsten om kwaliteit van zorg te meten (Bos et al., 2015). Deze handreiking legt op hoofdlijnen uit hoe partijen in de zorg vragenlijsten op een verantwoorde manier kunnen ontwikkelen en gebruiken om te komen tot valide en betrouwbare kwaliteitsinformatie. De criteria uit het Toetsingskader kwaliteitsstandaarden, informatiestandaarden en meetinstrumenten (Zorginstituut, 2015) zijn het uitgangspunt. Procedures en werkinstructies die eerder zijn ontwikkeld voor de Consumer Quality Index (CQ-index) methodiek waren de basis van de tools.

Zorginstituut Nederland biedt deze ondersteunende materialen aan om het meten, verzamelen en analyseren van kwaliteitsinformatie te standaardiseren. Hiermee wordt verdere verbetering van de kwaliteit van de gezondheidszorg in Nederland gestimuleerd en wordt ervoor gezorgd dat iedereen toegang heeft tot begrijpelijke en betrouwbare informatie over de kwaliteit van geleverde zorg. Onderzoekers van het Nederlands instituut voor onderzoek van de gezondheidszorg (NIVEL) en adviseurs van Zorginstituut Nederland hebben deze tool samengesteld .

Leeswijzer

Hoofdstuk 1 gaat in op verschillende voorbereidingsstappen en controles van de data. In hoofdstuk 2 wordt vervolgens bekeken of groepen vragen statistisch en inhoudelijk voldoende verwant zijn om een gemiddelde over die vragen te berekenen. Tot slot gaat hoofdstuk 3 over het uitvoeren van vergelijkende analyses voor afzonderlijke vragen of voor gemiddelden van groepen vragen.

Auteurs

Nivel: Dolf de Boer en Mattanja Triemstra
Zorginstituut Nederland: Laura Koopman en Marloes Zuidgeest

1. Voorbereiding

Patiëntervaringsvragenlijsten (Patient Reported Experience Measures (PREMs)[i] of Patient Reported Outcome Measures (PROMs[ii])) bieden de mogelijkheid om informatie te verzamelen waarbij de patiënt[1] centraal staat: zijn of haar mening of ervaring wordt gevraagd. Deze informatie kan gebruikt worden om te bep alen of zorgaanbieders[2] van elkaar verschillen in de ervaren kwaliteit van geleverde van zorg.

Met de verzamelde informatie moet zorgvuldig worden omgegaan bij het analyseren en presenteren. Fouten in databestanden die worden gebruikt bij analyses kunnen verregaande consequenties hebben op de resultaten van de door zorgaanbieders geleverde zorg. Daarom moet de verzamelde data eerst worden opgeschoond (zie 1.1) en gecontroleerd (zie 1.2) voordat deze geanalyseerd kan worden.

Voor het uitvoeren van deze controles wordt uitgegaan van een bestand met daarin alle benaderde patiënten en respondenten en de antwoorden op de vragen uit de ingevulde vragenlijsten. Doorgaans bestaat dit bestand uit enkele identificerende variabelen en uit een variabele voor iedere vraag. In Tabel 1 is een veel voorkomende lay-out van een dataset te zien. Hierin staan voorbeelden van identificerende variabelen zoals een geanonimiseerd uniek respondentnummer (idnummer), de naam of code van de zorgaanbieder, de naam/code van de meetorganisatie of een variabele die aangeeft met welke methode gegevens verzameld zijn (schriftelijk, online, interviews etc.). Per vraag zijn de antwoorden gecodeerd als cijfers, zoals bijvoorbeeld 1 'nooit', 2 'soms', 3 'meestal', 4 'altijd' en 9 'niet van toepassing'.

Tabel 1: Een voorbeeld van een deel van een databestand

Idnummer	Zorgaanbieder	Meetorganisatie	Methode	vraag1	vraag2	vraag3	vraag4
1234	ochtendgloren	A	Schriftelijk	3	2	3	2
1235	ochtendgloren	A	Online	4	4	3	9
1236	ochtendgloren	A	Schriftelijk	9	2	2	4
1237	avondrood	C	Online	3	3	4	4

1.1 Opschoning van het databestand

Voordat het databestand geschikt is om de ervaringen van patiënten te analyseren, dient deze opgeschoond te worden. De opschoning staat globaal weergegeven in Tabel 2.

Stap 1a Dubbelen verwijderen

De eerste stap in de opschoning is het bestand ontdubbelen en ontdoen van waarden waar niet mee gerekend mag worden. Hierbij kan het gaan om situaties waarin een ingevulde vragenlijst twee keer is ingescand, maar het komt ook voor dat respondenten in reactie op een herinnering de vragenlijst een tweede keer invullen. Voor beide situaties geldt dat slechts één record behouden kan blijven. Als er verschil zit in de versies moet de meest volledige bewaard worden. Als de twee versies hetzelfde zijn, wordt de eerste bewaard.

Stap 1b Missing maken

Bij het verwijderen van waarden waar niet mee gerekend mag worden, gaat het om irreële waarden (leeftijd >110; rapportcijfer > 10 etc.) of antwoorden als ("weet ik niet", "anders, namelijk:…", "niet van toepassing" etc.). Een eenvoudige oplossing is om deze waarden te verwijderen.

Totaalbestand

Na deze eerste stappen hebben we een 'Totaalbestand' dat bestaat uit alle personen die een vragenlijst hebben ontvangen (per post of e-mail) en de responsgegevens. Om te bepalen welke patiënten wel of niet gereageerd hebben en wat de hoogte is van de respons, dient een aantal aanvullende schonings-

1 Overal waar in deze tool de term "patiënt" wordt gebruikt, kan ook "cliënt" worden gelezen.

2 Zorgaanbieder is de algemene term die in deze tool wordt gebruikt om zorgaanbieders, verpleeghuizen, praktijken, ziekenhuizen, afdelingen aan te duiden.

stappen doorlopen te worden. Deze stappen hebben tot doel om alleen de antwoorden van patiënten te behouden die daadwerkelijk aangeschreven hadden mogen worden en die de vragenlijst voldoende en zelf hebben ingevuld. Deze stappen zijn eveneens weergegeven in Tabel 2 (zie stap 2a t/m 2c en stap 3a t/m 3c). In plaats van records direct uit het totaalbestand te verwijderen op grond van de schoningscriteria, kan voor elk schoningscriterium ook eerst een variabele worden aangemaakt. Deze variabele geeft dan aan of het schoningscriterium van toepassing is, zodat makkelijk te herleiden is of een respondent geëxcludeerd moet worden op grond van één of meerdere criteria. Op deze wijze blijft in één bestand zichtbaar welke records geschoond dienen te worden en waarom. Vlak voor de analyses kunnen deze records alsnog worden verwijderd, of kunnen de analyses alleen worden uitgevoerd op records waarop geen schonings-criterium van toepassing is.

Stap 2: Onterecht ontvangen vragenlijsten verwijderen.

In deze stap moeten vragenlijsten van bijvoorbeeld overledenen, onbestelbaar retour gestuurde vragen-lijsten en vragenlijsten ingevuld door patiënten die niet aan de in- of exclusiecriteria voldoen, verwijderd worden. Hierna blijft het netto aantal benaderde patiënten over.

Stap 3: Onvoldoende of niet juist ingevulde vragenlijsten verwijderen

Bijvoorbeeld als er lege vragenlijsten zijn ingestuurd, als onvoldoende vragen zijn beantwoord of de essentiële screenervragen nog leeg zijn, moeten deze vragenlijsten verwijderd worden. Ook hierbij is het mogelijk om de records niet gelijk te verwijderen maar door middel van een extra variabele aan te geven waarom deze gegevens niet gebruikt moeten worden.

Tabel 2. Schoningsstappen en berekening van het responspercentage, uit te voeren op het bestand met alle verstuurde vragenlijsten

Schoningsstap		Resultaat
1	Dubbelen verwijderen en missings (her)coderen	
1a	Dubbelen verwijderen	
1b	Waarden waar niet mee gerekend mag worden missing maken	
		Totaalbestand (na ontdubbeling)
2	Onterecht ontvangen vragenlijsten verwijderen, bijvoorbeeld:	
2a	Overledenen	
2b	Vragenlijsten die onbestelbaar retour kwamen	
2c	Respondenten die niet voldoen aan in- of exclusiecriteria*	
		Aantal netto benaderde patiënten
3	Onvoldoende of niet juist ingevulde vragenlijsten verwijderen, bijvoor-beeld:	
3a	Lege vragenlijsten	
3b	Als onvoldoende vragen zijn beantwoord**	
3c	Als essentiële vragen niet zijn ingevuld ***	
		Aantal vragenlijsten na schoning
	Responspercentage = Aantal vragenlijsten na schoning	
	Responspercentage = Aantal vragenlijsten na schoning * 100% Aantal netto benaderde patiënten	

* bijvoorbeeld als patiënten onder de 16 jaar voorkomen terwijl die buiten de doelgroep vallen

** bijvoorbeeld als minder dan de helft van de vragen die voor alle patiënten van toepassing zijn, zijn beantwoord

*** bijvoorbeeld vragenlijsten waarbij essentiële variabelen voor populatiekenmerken (zoals leeftijd, geslacht, zorgzwaarte, ook wel case-mix adjusters genoemd) missing zijn. Die vragenlijsten doen niet mee met vergelijkingen waarbij tussen zorgaan-bieders rekening wordt gehouden met deze kenmerken (case-mix correctie) (tenzij de missende informatie wordt geïmputeerd[iii] of wordt verkregen via een zorgaanbieder of zorgverzekeraar)

Het responspercentage wordt vervolgens berekend als het aantal vragenlijsten na schoning gedeeld door het aantal netto benaderde patiënten, vermenigvuldigd met 100 (%). Zie paragraaf 1.3 voor meer over responsanalyse.

Naast voornoemde stappen wordt in veel gevallen ook gecheckt of de antwoorden op screener- en vervolgvragen consistent zijn met elkaar (zie Box 1 voor een toelichting)

Box 1. Hoe om te gaan met screener- en vervolgvragen

Screenervragen zijn vragen waarin wordt bepaald of een vervolgvraag op de patiënten van toepassing is. Achter één van de mogelijke antwoorden op de screenervraag staat dan de instructie dat de patiënt vervolgvragen niet hoeft in te vullen. Een veelgebruikte vuistregel is dat de screenervraag leidend is. Als uit de screenervraag volgt dat vervolgvragen niet van toepassing zijn, worden deze als missing gecodeerd als ze toch zijn ingevuld. Vanuit dezelfde gedachte worden antwoorden op vervolgvragen vaak ook missend gemaakt als de screenervraag niet ingevuld is. Het is van belang om bij screenervragen per situatie te bekijken of de beschreven vuistregels passend zijn.

Een voorbeeld van een screenervraag (Bron: CQ-index Fysiotherapie, versie 2.3)

43. Bent u voor uw klachten door verschillende fysiotherapeuten behandeld?

☐ Nee → Indien nee, ga door naar vraag 45

☐ Ja

Tips

- Gebruik voor de opschoning het 'totaalbestand'. In dit bestand staat per patiënt genoteerd wie wel (respondent) en wie niet (non-respondent) de vragenlijst heeft ingevuld.
- Het totaalbestand kan gebruikt worden om de representativiteit te bepalen van de groep patiënten die de vragenlijst hebben ingevuld (zie 1.4.1).
- Bepaal met het geschoonde totaalbestand de hoogte van het responspercentage (zie 1.3).

1.2 Controles op coderingsfouten

Een coderingsfout betreft de situatie dat een waarde voor een variabele niet de afgesproken betekenis heeft. De afspraak kan bijvoorbeeld zijn dat geslacht wordt gecodeerd als vrouw=0 en man=1. Als een deel of geheel van de dataset anders is gecodeerd (bijvoorbeeld: vrouw=1 en man=0, of vrouw=1 en man=2), dan is dat een coderingsfout omdat dit niet voldoet aan de afgesproken betekenis.

1.2.1 Controle op coderingsfouten betreffende de gehele dataset:

Check of antwoorden op vragen logisch zijn. Bijvoorbeeld bij een vraag als "Was uw zorgverlener beleefd tegen u?" is bekend dat de meerderheid van de patiënten daar (heel) positief op antwoordt. Wanneer in een dataset blijkt dat de overgrote meerderheid hier juist negatief op antwoordt kan dat ook een aanwijzing zijn voor een coderingsfout.

Check bij vragen die negatief geformuleerd zijn, zoals "Heeft u tegenstrijdige informatie ontvangen?" of antwoorden daar ook goed gecodeerd zijn.

1.2.2 Controle op coderingsfouten van een deel van de dataset.

Vergelijk per vraag de gemiddelde, mediaan, frequentieverdelingen, en het percentage missende waarden tussen:

- meetorganisaties
- methoden van dataverzameling (zoals schriftelijk en online)
- ICT-registratiesystemen (indien bekend)
- Zorgaanbieders

Opmerkelijke verschillen zoals hogere gemiddelde waarden, afwijkende verdelingen of hogere percentages missende waarden kunnen duiden op een coderingsfout. Deze fouten dienen onderzocht te worden en als het coderingsfouten betreffen moeten deze worden opgelost.

1.2.3 Coderingsfouten oplossen

Indien duidelijk is op welk deel van de data een mogelijke coderingsfout betrekking heeft, dient deze gecheckt te worden bij een meetorganisatie, ICT registratiesysteem of anderszins. Correctie kan plaatsvinden met een hercodering. Idealiter wordt de hercodering uitgevoerd door de partij die de coderingsfout heeft gemaakt. Soms leidt dat ertoe dat die partij een nieuwe datalevering moet doen aan een centrale database. Indien daar geen tijd voor is kan een hercodering ook bij de centrale database worden toegepast, uiteraard na zorgvuldige afstemming met de partij waar de fout is ontstaan en gevolgd door strenge controles op de hercodering.

Indien een mogelijke coderingsfout heeft plaatsgevonden bij één zorgaanbieder is het de vraag of het wel gaat om een coderingsfout, of dat het gaat om een afwijkende zorgaanbieder. Afhankelijk van de aard en omvang van de afwijking en de checkmogelijkheden wordt besloten dergelijke afwijkingen nader te onderzoeken. Aanbevolen wordt om forse afwijkingen, gebaseerd op veel patiënten, altijd nader te onderzoeken

1.3 Responsanalyse

Zoals in tabel 2 staat beschreven kan, door de schoningsstappen in 1.1 te doorlopen, het responspercentage berekend worden door het aantal ingevulde vragenlijsten na schoning te delen door het aantal netto benaderde patiënten. Gemiddeld is een responspercentage van 50% al heel mooi bij metingen met schriftelijke vragenlijsten. Veel is meer dan 70%, weinig is minder dan 20%. Maar bij online vragenlijsten is 20% al hoog. Check tussen zorgaanbieders de hoogte van respons. Daar waar de respons van een zorgaanbieder erg laag is in vergelijking met de andere zorgaanbieders (bijvoorbeeld 10% versus 70%) kan dit aanleiding zijn voor nader onderzoek naar het verloop van het dataverzamelingsproces. Mogelijke verklaringen zijn dat vragenlijsten niet naar alle patiënten verstuurd zijn, of dat het technisch niet mogelijk was om voor bepaalde patiënten in te loggen om een vragenlijst online in te vullen.

Voor de geldigheid van het onderzoek is het van belang om te weten of de groep patiënten die gereageerd hebben een goede afspiegeling vormt van de hele doelgroep.

1.4 Representativiteit: vormen respondenten een goede afspiegeling van alle patiënten?

De representativiteit kan worden vastgesteld door de groep patiënten die gereageerd hebben (respondenten) te vergelijken met alle benaderde patiënten, die in eerste instantie de vragenlijst ontvingen (steekproef[3]). De representativiteit wordt getoetst door de verdeling van leeftijd en geslacht te vergelijken tussen de respondenten en de netto benaderde patiënten (zie Tabel 2). Er is sprake van een representatieve respons als de responspercentages per leeftijdscategorie en sekse ongeveer gelijk zijn aan die van de netto benaderde patiënten.

Wanneer tijdens de ontwikkeling van de patiëntervaringsvragenlijst geen informatie beschikbaar is over de kenmerken van patiënten (in de steekproef of populatie), dan kunnen de kenmerken van patiënten in de responsgroep vergeleken worden met gegevens uit andere bronnen. Wanneer een meting representatief moet zijn voor de algemene bevolking, kunnen de respondenten bijvoorbeeld worden vergeleken met gegevens van het CBS over de bevolkingssamenstelling. Bij specifieke patiëntengroepen kan het zijn dat er landelijke cijfers zijn over de samenstelling van die groepen, bijvoorbeeld in jaarrapportages van de beroepsgroep. Ook als er wel cijfers beschikbaar zijn van kenmerken van patiënten (steekproef of populatie) kan via deze vergelijking bekeken worden of de groep patiënten representatief is voor de algemene danwel specifieke bevolking. Dit is altijd nuttig om te doen omdat het aanvullend inzicht geeft in de representativiteit van de respondenten.

3 In sommige metingen wordt gewerkt met een steekproef, in andere metingen wordt de hele populatie van een zorgaanbieder benaderd voor deelname. Waar we in deze tool spreken over een steekproef kan ook de gehele populatie worden gelezen.

2. Groeperen van vragen: Factoranalyse, betrouwbaarheidsanalyse en IRT

Het vergelijken van patiëntervaringen tussen zorgaanbieders kan op basis van afzonderlijke vragen uit de vragenlijst, maar ook op basis van gemiddelden over groepen vragen. Deze laatste methode leidt doorgaans tot betrouwbaardere resultaten omdat bij het berekenen van gemiddelden, meetfouten worden uitgemiddeld. Om te beoordelen of, en zo ja welke groepen vragen in aanmerking komen voor het berekenen van gemiddelden, worden vaak factoranalyses (2.2)[iv] en betrouwbaarheidsanalyses (2.3) v uitgevoerd. Daarnaast worden steeds vaker technieken toegepast die gebaseerd zijn op Item Respons Theorie (IRT)[vi]. Tijdens de ontwikkeling van een vragenlijst wordt vaak intensief aandacht besteed aan dit soort technieken om de eigenschappen van de vragen(lijst) goed in beeld te brengen en te valideren. Bij het gebruik van een reeds ontwikkelde vragenlijst is het vaak nuttig om deze technieken ook (deels) toe te passen om te checken of de vragenlijst dezelfde eigenschappen laat zien in de nieuwe dataset.

Voor ieder van de genoemde technieken geldt dat er vele tekstboeken beschikbaar zijn waar ze in al hun varianten en verschijningsvormen worden beschreven. In deze sectie volstaan we dan ook met het noemen van enkele belangrijke overwegingen en keuzes en verwijzen we naar tekstboeken voor meer informatie.

2.1. Eigenschappen van de vragen inspecteren

Om een indruk te krijgen van de meeteigenschappen van de vragenlijst, is het aan te bevelen om de eigenschappen van de vragen (variabelen) te inspecteren voordat met de analyses wordt begonnen. Dit is iets anders dan schonen: het gaat nu niet om dingen die niet kloppen of onvoldoende zijn ingevuld, maar om de validiteit en verdeling van antwoorden op de vragen. Denk hierbij aan:

- Scheefheid van de antwoorden op een vraag. Stel bijvoorbeeld dat meer dan 90% van de respondenten hetzelfde antwoordt op een vraag. Antwoorden die zo scheef verdeeld zijn, zullen naar verwachting niet of nauwelijks verschillen laten zien tussen patiëntengroepen of zorgaanbieders.
- Missende waarden per vraag. Wanneer een vraag heel vaak niet wordt beantwoord, of wordt beantwoord met “niet van toepassing” of “weet ik niet”, dan kan dat een aanwijzing zijn dat de vraag niet goed begrepen wordt. Als vuistregel kunnen vragen met meer dan 5% missende waarden nader worden bekeken.
- Samenhang tussen vragen. Indien vragen samenhangen met een Pearson’s correlatie coëfficiënt van $r \geq 0{,}70$ is er aanzienlijke overlap in de informatie die beide vragen verschaffen. Dit kan aanleiding zijn om één van beide vragen niet verder te analyseren.

Deze criteria geven een indruk van de geschiktheid van de vragen voor de analyses. Bij reguliere metingenvi komt het vaak voor dat in een eerder stadium al besluiten zijn genomen over de samenstelling van de vragenlijst en de te analyseren vragen. Dan kan deze stap eventueel worden overgeslagen. Toch is het altijd zinnig om bij (vervolg)metingen vast te stellen of de vragenlijst opnieuw dezelfde meeteigenschappen laat zien.

2.2. Factoranalyse

Factoranalyse is een statistische techniek waarmee voor een groot aantal vragen een beperkter aantal achterliggende variabelen (factoren) kan worden geïdentificeerd. Verschillende vragen over bejegening (luisterde de arts goed, had de arts voldoende tijd, …. etc.) vormen bijvoorbeeld vaak samen een factor of ‘schaal’. Met factoranalyse wordt statistisch bekeken welke vragen voldoende samenhangen om een factor te vormen. Van belang is echter dat de vragen die bij een factor horen niet alleen statistisch samenhangen, maar ook inhoudelijk verwant zijn. Of vragen inhoudelijk verwant zijn, wordt niet statistisch getoetst, maar is ter beoordeling aan degene die de factoranalyse uitvoert en/of inhoudsdeskundigen zoals bijvoorbeeld patiënten zelf, of professionals.,

Factoranalyse wordt vaak geassocieerd met de validiteit van een vragenlijst, hierbij gaat het om de kwestie of de vragenlijst daadwerkelijk meet wat je wil meten. Meer specifiek is factoranalyse een manier om zicht te krijgen op de structurele validiteit en daarmee wordt factoranalyse ook gezien als een vorm van constructvaliditeit (Mokkink et al., 2010). Het gaat hierbij om welke vragen samen dimensies (factoren of schalen) vormen en of dat consistent is met de constructen (onderwerpen) die de vragenlijst beoogd te meten.

2.2.1. Factoranalyse: exploratief of confirmatief?

Globaal zijn er twee soorten factoranalyse: exploratieve en confirmatieve factoranalyse. Bij een exploratieve factoranalyse wordt een groep vragen geanalyseerd, zonder vooraf te specificeren welke groepen van vragen samen factoren vormen. Vaak komen er factoroplossingen uit die ook inhoudelijk aannemelijk zijn, dat wil zeggen dat vragen die samen een factor vormen ook inhoudelijk verwant zijn. Het gebeurt echter ook regelmatig dat een vraag bij een factor wordt ondergebracht die daar inhoudelijk helemaal niet bij past. Het blijft dus van belang om te beoordelen of de oplossingen van een exploratieve factoranalyse ook inhoudelijk aannemelijk zijn.

Een andere vorm van factoranalyse is confirmatieve factoranalyse. Bij een confirmatieve factoranalyse wordt vooraf gespecificeerd welke vragen naar verwachting samen factoren vormen en in hoeverre verschillende factoren naar verwachting samen hangen. Vervolgens wordt met de confirmatieve factoranalyse getoetst of de verwachte factorstructuur past bij de data. De verwachte factorstructuur voor een confirmatieve factoranalyse wordt vaak gebaseerd op inhoudelijke overwegingen (welke vragen lijken op inhoudelijke gronden bij elkaar te horen?), al dan niet aangevuld met de bevindingen van een exploratieve factoranalyse of bevindingen uit eerder onderzoek.

Meer informatie over exploratieve en confirmatieve factoranalyse is te vinden in tal van tekstboeken, zoals bijvoorbeeld Dennis (2006) en Thompson (2004).

2.3. Betrouwbaarheidsanalyse

Betrouwbaarheid is de mate waarin een meting vrij is van meetfouten (Mokkink et al., 2010). Doorgaans is een gemiddelde of somscore over een aantal vragen (thema, factor, dimensie, composiet of schaalvii) betrouwbaarder dan de score op de afzonderlijke vragen. In het resterende deel van dit document spreken we van een thema[v].

2.3.1. Betrouwbaarheid van thema's vaststellen

De meest gangbare manier om vast te stellen of een thema werkelijk betrouwbaar is, is met behulp van Cronbach's alfa coëfficiënt, een getal tussen 0 en 1. Hierbij wordt een waarde van hoger dan 0,7 vaak als voldoende betrouwbaar beschouwd. Indien Cronbach's alfa veel lager is dan 0,70, dan kan dit soms worden opgelost door één van de vragen te laten vervallen om de betrouwbaarheid te vergroten. Helpt dit niet, dan kunnen ook één à twee vragen worden gekozen die het thema vertegenwoordigen. Die vragen worden dan als afzonderlijk geanalyseerd en gepresenteerd.

Wanneer de onderliggende vragen van een thema niet op een consistente manier zijn gesteld, vraagt dit de nodige aandacht bij de bepaling van de betrouwbaarheid. Hierbij kan sprake zijn in verschillende situaties.

'Andersom' gestelde vragen.
Bij de meeste vragenlijsten geldt dat vragen overwegend in dezelfde 'richting' zijn gesteld, bijvoorbeeld dat bevestigende antwoorden (zoals 'ja' of 'altijd') op positieve ervaringen duiden. Wanneer er vragen voorkomen die juist op een andere manier gesteld zijn, dienen de antwoorden op die vragen te worden omgecodeerd. Een voorbeeld is de vraag 'hoe vaak had u overlast van andere patiënten op de kamer?' waarbij het antwoord 'altijd' op een negatieve ervaring duidt. Als dergelijke 'andersom'-vragen onderdeel uitmaken van een schaal, moeten zij vóór het bepalen van het schaalgemiddelde zodanig gecodeerd of 'omgepoold' worden dat de meest positieve antwoordcategorie hetzelfde cijfer krijgt als de meest

positieve antwoordcategorie voor de overige vragen. Bijvoorbeeld: bij een vraag met vier antwoordcategorieën wordt 1 dan 4; 2 wordt 3; 3 wordt 2; en 4 wordt 1.

Verschillende aantallen antwoordcategorieën.
Het kan zijn dat een schaal bestaat uit een combinatie van vragen met twee antwoordcategorieën (bijvoorbeeld 'ja' en 'nee') en vragen met vier antwoordcategorieën (bijvoorbeeld 'nooit', 'soms', 'meestal', 'altijd'). In dat geval moeten de vragen met twee antwoordcategorieën vóór het bepalen van het themagemiddelde worden gehercodeerd, zodanig dat de meest positieve antwoorden van beide vragen dezelfde waarde krijgen, én de meest negatieve antwoorden van beide vragen dezelfde waarde krijgen. Bijvoorbeeld: het antwoord 'ja' krijgt hetzelfde cijfer als het antwoord 'altijd' (de waarde '4') en het antwoord 'nee' krijgt hetzelfde cijfer als het antwoord 'nooit' (de waarde '1').

2.3.2. Gemiddelden berekenen over de vragen uit een thema

Het themagemiddelde per respondent wordt meestal berekend als het (ongewogen) gemiddelde van de scores op de afzonderlijke vragen. In de meeste gevallen is er een minimum voor het aantal te beantwoorden vragen van een thema. Bijvoorbeeld de voorwaarde dat alleen respondenten die meer dan de helft van de vragen per thema hebben beantwoord, worden meegenomen in de berekening van themagemiddelden. Deze berekening wordt geïllustreerd aan de hand van onderstaand voorbeeld, waarbij de scores van vijf respondenten op drie vragen (die tot één schaal behoren) worden gebruikt voor het berekenen van een themagemiddelde (zie de laatste kolom).

Tabel 3: Voorbeeld berekening gemiddelde per schaal

Vraag 1	Vraag 2	Vraag 3	Thema	Gemiddelde
Respondent 1	2	4	4	10/3 = 3,333
Respondent 2	3	4	4	11/3 = 3,666
Respondent 3	2	3	Missing	5/2 = 2,5
Respondent 4	4	4	3	11/3 = 3,666
Totaal				**3,292**

Een aandachtspunt bij het berekenen van themagemiddelden zoals in tabel 3, is dat er bij missende waarden impliciet vanuit wordt gegaan dat deze gelijk zijn aan het gemiddelde van de beschikbare vragen. In de tabel is dit te zien bij respondent 3, die op grond van twee beschikbare waarden de score 2,5 krijgt, alsof de missende derde waarde ook 2,5 zou bedragen. Hoewel dat heel vaak op deze manier wordt gedaan, zijn er ook alternatieve oplossingen. Bijvoorbeeld door voor de ingevulde vragen te bekijken of die persoon daarvoor boven of juist onder het gemiddelde scoort, om voor de ontbrekende waarden voor die persoon een vergelijkbare onder- of bovengemiddelde score in te vullen. Een dergelijke oplossing voorkomt dat missings op vragen die veel positiever of negatiever worden beantwoord dan de andere vragen in het thema, het themagemiddelde vertekenen.

2.4. Item Respons Theorie

Met analyses gebaseerd op item respons theorie (IRT) is het mogelijk om vragen die hetzelfde construct meten, te rangschikken op een schaal naar de "moeilijkheid" van die vragen. De vragen krijgen dan allemaal een locatie op de schaal waarop het construct wordt gemeten. Hardlopen is bijvoorbeeld aanzienlijk moeilijker dan wandelen, dus een vraag of iemand 10 minuten kan hardlopen zal een hogere locatie krijgen op de schaal dan de vraag of iemand 10 minuten kan wandelen. Een groot voordeel van item respons analyses is dat niet alleen de vragen een locatie krijgen op de schaal, ook personen krijgen op basis van hun antwoorden een locatie op dezelfde schaal. Op grond van de score van een individu is dan direct duidelijk welke vragen voor dat individu moeilijk(er) zijn en welke makkelijk(er) zijn. Item respons modellen zijn heel populair bij vragenlijsten waarin het niveau van fysiek, mentaal of sociaal functioneren wordt gemeten (PROMs) en bij intelligentietesten of vaardigheidstesten. Voor PREMs, zoals de CQ-index, worden item respons modellen veel minder gebruikt. Over Item Respons Theorie en de bijbehorende analyses is meer te lezen in verschillende boeken, waaronder van der Linden et al., (1996) en Embretson en Reise (2000).

3. Vergelijkende analyses uitvoeren

Voor het uitvoeren van vergelijkende analyses tussen zorgaanbieders wordt in deze tool uitgegaan van multi-level analyse (zie bijlage I voor een toelichting). Deze methode houdt rekening met de structuur van de data: ervaringen van patiënten zijn geclusterd binnen zorgzorgaanbieders en zijn daarmee niet onafhankelijk van elkaar. Schattingen per zorgaanbieder uit een multi-level model staan onder verschillende namen bekend: Emperical Bayes schattingen, posterior means of shrinkage estimators. In deze tool spreken we van Emperical Bayes (EB) schattingen

Andere methoden die gebruikt kunnen worden voor het vergelijken van patiëntervaringen tussen zorgaanbieders zijn onder meer directe en indirecte standaardisatie (Curtin en Klein, 1995) of subgroepanalyses (aparte vergelijkingen voor ieder niveau van de controle variabele). Deze andere methoden worden in deze tool niet uitgewerkt.

3.1. Aantallen zorgaanbieders en respondenten per zorgaanbieder

3.1.1. Aantal zorgaanbieders

Tijdens het uitvoeren van vergelijkende analyses is het van groot belang om te beschikken over voldoende zorgaanbieders in de dataset die het liefst ook representatief zijn voor alle zorgaanbieders in Nederland. Een bekende vuistregel uit de multi-level literatuur is het aantal van 30 zorgaanbieders (Hox, 1998). In de praktijk wordt ook wel gewerkt met datasets van 20 zorgaanbieders. Naarmate het aantal beschikbare zorgaanbieders lager wordt, is het van groter belang dat de zorgaanbieders in de dataset (vrijwel) alle zorgaanbieders in Nederland zijn zodat in ieder geval duidelijk is dat die zorgaanbieders representatief zijn.

3.1.2. Aantal waarnemingen per zorgaanbieder

Meestal is uit eerdere (pilot-)metingen bekend hoeveel patiënten/respondenten per zorgaanbieder nodig zijn om verschillen tussen zorgaanbieders significant aan te tonen. Dit is bijvoorbeeld terug te vinden in een werkinstructie of een onderzoeksrapport over de vragenlijst. Als het beoogde aantal patiënten voor (veel) zorgaanbieders niet gehaald is, zullen de resultaten weinig onderscheid laten zien tussen zorgaanbieders.

Indien uit een eerdere meting niet bekend is hoeveel patiënten nodig zijn per zorgaanbieder om verschillen tussen zorgaanbieders aan te kunnen tonen, is het aan te raden over minimaal enkele tientallen patiënten per zorgaanbieder te beschikken. Op grond van deze resultaten kan berekend worden hoeveel patiënten in een toekomstige meting nodig zijn om verschillen tussen zorgaanbieders betrouwbaar te kunnen meten (zie 3.5.1).

3.2. Specificeren variabelen

Verschillen in populatiekenmerken (onafhankelijke variabelen zoals leeftijd, geslacht en zorgzwaarte) tussen zorgaanbieders hebben (mogelijk) invloed op de resultaten uit de vergelijkende analyses. Daarom wordt in de regel gecorrigeerd voor deze populatiekenmerken of case-mix adjusters, die bij voorkeur worden gecentreerd voordat ze meegenomen worden in de multi-level analyses. Het is raadzaam om case-mix-variabelen te centreren (Enders & Tofighi, 2007). Centreren gebeurt door voor de betreffende variabele het gemiddelde van alle scores af te trekken. Voor een case-mix adjuster waarbij het noodzakelijk is dummyvariabelen aan de dataset toe te voegen (bijvoorbeeld leeftijdscategorieën), is een pragmatische oplossing om de meest voorkomende antwoordcategorie als referentiegroep te nemen. Door te centreren liggen de voor case-mix gecorrigeerde schattingen per zorgaanbieders dichter bij hun ruwe gemiddelden. Daardoor benaderen de resultaten beter de werkelijkheid, wat de interpretatie vergemakkelijkt.

Bij afhankelijke variabelen (vragen of thema's) is met name het type model van belang. Er wordt gebruik gemaakt van lineaire, ordinale of logistische modellen. Dit is afhankelijk van het aantal categorieën of

waarden die de variabele kan aannemen. Tabel 4 bevat enkele vuistregels voor de keuze van een lineair, ordinaal of logistisch multi-level model.

Tabel 4. Aantal categorieën bij afhankelijke variabelen en type multilevel model*

Aantal categorieën	Type multi-level model	Referentie
≥ 7	Lineair	(Tabachnik & Fidell, 2001, p7)
3-6	Ordinaal	
2	Logistisch	

*Zie voor meer informatie over verschillende typen modellen en hun eigenschappen: De Boer en Van der Hoek (2015)

3.3. Aantal niveaus in het multi-level model

De multi-level analyse kent minimaal twee niveaus: die van de patiënt – het individu – en het niveau waarop de resultaten worden vergeleken, oftewel de eenheid van analyse (de zorgaanbieder, het ziekenhuis, de afdeling, locatie of praktijk etc.). Soms zijn er meer dan twee niveaus beschikbaar in de data, bijvoorbeeld de patiënt, de zorgverlener en de zorgaanbieder. Over het algemeen is het methodisch beter om een extra niveau mee te nemen in de analyse als dat niveau in de data beschikbaar is. In de praktijk zijn dergelijke extra niveaus niet altijd (volledig) beschikbaar in de dataset.

3.4. Schattingen per zorgaanbieder

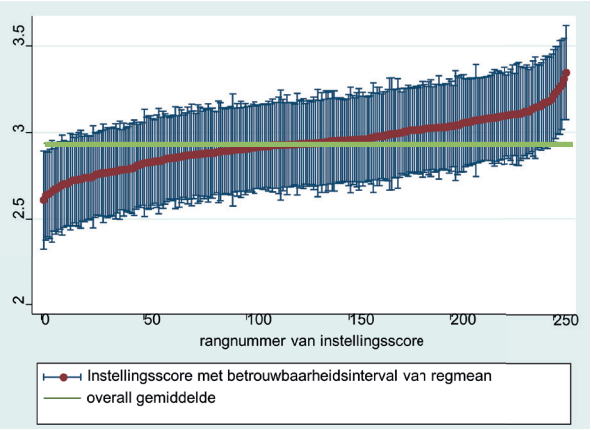
Schattingen van themagemiddelden of item-scores per zorgaanbieder met een multi-level model liggen tussen het ruwe gemiddelde per zorgaanbieder en het gemiddelde van alle zorgaanbieders in. Dit type schattingen staat onder verschillende namen bekend, waaronder shrinkage schattingen, empirical Bayes schattingen en posterior means (zie bijlage 1 voor een toelichting).

Schattingen per zorgaanbieder (of enige andere analyse-eenheid) mogen alleen gedaan worden indien het multi-level model significant verschilt van een model zonder niveaus, bijvoorbeeld als dit blijkt uit een likelihood ratio test. Indien dit niet het geval is, dan zijn er geen verschillen tussen de zorgaanbieders. Als dan toch schattingen worden gegenereerd, dan zullen deze geen betrouwbare verschillen tussen zorgaanbieders laten zien. De geschatte scores per zorgaanbieder kunnen worden gepresenteerd voor alle zorgaanbieders (3.4.1) of per zorgaanbieder (3.4.2).

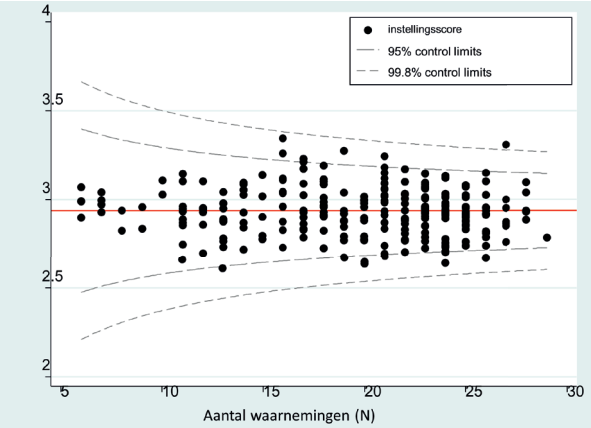
3.4.1. Presentatie van alle zorgaanbieders

Veel voorkomende presentatiewijzen voor het presenteren van alle schattingen in één grafiek zijn wormgrafieken of 'caterpillars' (zie: figuur 1) (Goldstein & Healy, 1995) en funnelplots (zie: figuur 2) (Spiegelhalter, 2005). Ze helpen om een totaalbeeld te vormen van de resultaten. Mochten hierbij vreemde patronen zichtbaar worden, dan kan dat aanleiding zijn voor aanvullende controles van data en analyses.

Figuur 1: Een voorbeeld van een wormgrafiek (caterpillar plot) van de scores per zorgaanbieder en hun betrouwbaarheidsinterval. De y-as geeft de scores (regressiegemiddelden, regmean) weer.



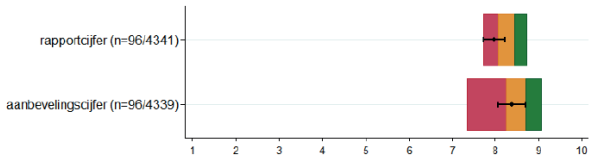
Figuur 2: Een voorbeeld van een funnelplot met de prestatiescore op de y-as en het aantal waarnemingen per zorgaanbieder op de x-as. Scores buiten de 95%-limits zijn opmerkelijk en scores buiten de 99,8% limits zijn zeer opmerkelijk.



3.4.2. Presentatie per zorgaanbieder

Een presentatiewijze die meer is gericht op het presenteren van schattingen voor één zorgaanbieder is te vinden in figuur 3. Deze presentatiewijze is ontwikkeld door het Picker Institute en laat zien hoe een zorgaanbieder scoort ten opzichte van de laagste 20%, de middelste 60% en de hoogste 20% van scores per zorgaanbieder. Deze presentatiewijze wordt in Nederland ook steeds vaker gebruikt om scores terug te koppelen aan individuele zorgaanbieders.

Figuur 3: Een voorbeeld van het presenteren van een score van een zorgaanbieder met betrouwbaarheidsinterval ten opzichte van de laagste 20% (rood), de middelste 60% (oranje) en de hoogste 20% (groen) van de scores per zorgaanbieder.



Een verder vereenvoudigde presentatiewijze is een indeling in groepen, die bijvoorbeeld ‘gemiddeld’, ‘bovengemiddeld’ of ‘benedengemiddeld’ scoren, zoals bij de CQ-index methodiek gebeurde middels het toekennen van sterren. Door hierbij rekening te houden met de onzekerheidsmarges van de scores per zorgaanbieder kunnen sterren worden toegekend, zodanig dat zorgaanbieders die als benedengemiddeld worden geclassificeerd, significant verschillen van zorgaanbieders die als bovengemiddeld worden geclassificeerd (zie box 4).

Box 4. Vereenvoudigde presentatie van resultaten door het berekenen van vergelijkingsintervallen en het toekennen van sterren aan zorgaanbieders: een voorbeeld

Bij de CQ-index methode was het gebruikelijk om scores per zorgaanbieder te classificeren als benedengemiddeld (1 ster), gemiddeld (2 sterren) en bovengemiddeld (3 sterren). Hiertoe werd voor iedere score van een zorgaanbieder een vergelijkingsinterval berekend als het gecorrigeerde gemiddelde +/- de standaardfout * 1,39 (zie voor een toelichting: Goldstein & Healy, 1995). Als de vergelijkingsintervallen tussen twee zorgaanbieders niet overlappen is het verschil tussen beide zorgaanbieders significant (p<.05).

De sterren werden als volgt toegekend:

1 ster: als de bovengrens onder het gemiddelde over alle zorgaanbieders ligt (de zorgaanbieders linksonder in figuur 1);

2 sterren: als het vergelijkingsinterval overlapt met het gemiddelde over alle zorgaanbieders (de zorgaanbieders in het midden van figuur 1).

3 sterren: als de ondergrens van het vergelijkingsinterval geheel boven het gemiddelde over alle zorgaanbieders ligt (de zorgaanbieders rechts boven in figuur 1);

Zorgaanbieders met 1 ster verschillen dus significant van zorgaanbieders met 3 sterren.

3.5. Betrouwbaarheid van scores per zorgaanbieder en schaduwanalyse

3.5.1. Betrouwbaarheid van de scores per zorgaanbieder

De betrouwbaarheid van geschatte waarden per zorgaanbieder wordt groter naarmate de verschillen tussen zorgaanbieders groter zijn. Een eerste indicatie van de omvang van de verschillen tussen zorgaanbieders is de Intra Class Correlation Coefficiënt [ix] (ICC). Dit is een getal tussen nul en één en geeft weer hoeveel procent van de variantie in patiëntervaringen is toe te schrijven aan verschillen tussen zorgaanbieders (Snijders & Bosker, 1999). Naarmate de ICC hoger is, zijn er minder waarnemingen per zorgaanbieder nodig om de scores per zorgaanbieder betrouwbaar te kunnen schatten. Een punt van aandacht is dat relatief hoge ICC's (>0,1) soms ook kunnen duiden op een coderingsfout van bijvoorbeeld een meetorganisatie. Een hoge ICC kan dus aanleiding zijn voor een aanvullende controle.

Een tweede belangrijke controle betreft het berekenen van de zogeheten reliabilities. De reliability[x] per zorgaanbieder is een maat voor de precisie van geschatte waarde per zorgaanbieder en is – net als de ICC – een getal tussen nul en één. De reliability per zorgaanbieder hangt af van de grootte van de verschillen tussen zorgaanbieders en van het aantal waarnemingen voor een zorgaanbieder. Lage reliabilities per zorgaanbieder (<0,70) suggereren dat er te weinig waarnemingen per zorgaanbieder zijn om scores per zorgaanbieder betrouwbaar te kunnen schatten. Voor toekomstige metingen is het dan informatief om uit te rekenen hoe hoog het aantal waarnemingen per zorgaanbieder had moeten zijn om wel reliabilities van 0,70 te behalen. Dit is van belang om terug te koppelen aan de opdrachtgever van de analyses en/of andere betrokken partijen. Voor een lineair model is de reliability λ bij benadering:

$$\lambda_j = \tau^2_o / (\tau^2_o + \sigma^2/n_j) \quad (1)$$

waarbij τ^2_o staat voor de variantie tussen organisatorische eenheden, σ^2 staat voor de variantie binnen organisatorische eenheden en nj het aantal waarnemingen weergeeft binnen organisatorische eenheid *j* (Snijders & Bosker, 1999). Voor een logistisch en voor een ordinaal model is de reliability λ bij benadering:

$$\lambda_j = \tau^2_o / (\tau^2_o + (\pi^2/3)/n_j) \quad (2)$$

waarbij π staat voor het getal pi. Bij formules (1) en (2) geldt dat deze de reliability benaderen omdat de reliability vooral van belang is voor het model met case-mix adjusters terwijl deze formules in principe bedoeld zijn voor modellen zonder case-mix adjusters. Deze reliabilities geven inzicht in de precisie waarmee scores per zorgaanbieder kunnen worden berekend en – daarmee samenhangend – hoe

informatief het is om op basis van die scores vergelijkingen te trekken tussen zorgaanbieders. Ook in de internationale literatuur wordt de reliability veelvuldig gebruikt als maat voor de precisie van scores per zorgaanbieder (Fung et al., 2010; Hargraves, Hays, & Cleary, 2003; Keller et al., 2005; Solomon, Hays, Zaslavsky, Ding, & Cleary, 2005).

Tip

Controleer resultaten zorgvuldig. Dit kan via een schaduwanalyse. Dit betreft het vergelijken van resultaten voor de centreringen, de modellen, de scores per zorgaanbieder en de standaardfouten tussen twee onafhankelijke personen. Inconsistenties tussen hoofdanalyse en schaduwanalyse vereisen dan nadere controle.

Bronnen en verder lezen

- Arling, G., Lewis, T., Kane, R. L., Mueller, C., & Flood, S. (2007). Improving quality assessment through multilevel modeling: the case of nursing home compare. Health Serv.Res., 42(3 Pt 1), 1177-1199.
- Curtin LR, Klein RJ. (1995). Direct standardization (age-adjusted death rates). Healthy People 2000 Stat Notes. 1995 Mar;(6):1-10.
- Damman, O. C., Stubbe, J. H., Hendriks, M., Arah, O. A., Spreeuwenberg, P., Delnoij, D. M., & Groenewegen, P. P. (2009). Using multilevel modeling to assess case-mix adjusters in consumer experience surveys in health care. Med.Care, 47(4), 496-503.
- Dennis, C. (2006). The Essentials of Factor Analysis (3rd ed.). Continuum International.
- de Boer, D., & van der Hoek, L. (2015). Het vergelijken van patiëntervaringen tussen zorgaanbieders. Meerwaarde van ordinale modellen voor indicatoren gebaseerd op één vraag met een beperkt aantal antwoordcategorieën.
- De Boer, D., van der Hoek, L., Delnoij, D. M., & Groenewegen, P. (2010). Kleine zorgaanbieders in multilevel vergelijkende analyses. De CQI Verpleging, Verzorging en Thuiszorg. Retrieved from
- Embretson, Susan E.; Reise, Steven P. (2000). Item Response Theory for Psychologists. Psychology Press.
- Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. Psychol.Methods, 12(2), 121-138.
- Fung, V., Schmittdiel, J. A., Fireman, B., Meer, A., Thomas, S., Smider, N., . . . Selby, J. V. (2010). Meaningful variation in performance: a systematic literature review. Med Care, 48(2), 140-148.
- Goldstein, H., & Healy, M. J. R. (1995). The Graphical Presentation of a Collection of Means. Journal of the Royal Statistcial Society.Seris A (Statistics in Society), 158(1), 175-177.
- Goldstein, H., & Spiegelhalter, D. J. (1996). League Tables and their Limitations: Statistical Issues in Comparisons of Institutional Performance. Journal of the Royal Statistcial Society.Seris A (Statistics in Society), 159(3), 385-443.
- Hargraves, J. L., Hays, R. D., & Cleary, P. D. (2003). Psychometric properties of the Consumer Assessment of Health Plans Study (CAHPS) 2.0 adult core survey. Health Serv.Res., 38(6 Pt 1), 1509-1527.
- Hox, J. (1998). Multilevel Modeling: When and Why. In I. Bahlderjan, R. Mathar, & M. Schader (Eds.), Classification, data analysis, and data highways. New York: Springer Verlag.
- Keller, S., O'Malley, A. J., Hays, R. D., Matthew, R. A., Zaslavsky, A. M., Hepner, K. A., & Cleary, P. D. (2005). Methods used to streamline the CAHPS Hospital Survey. Health Serv Res, 40(6 Pt 2), 2057-2077.
- Van der Linden, Wim J.; Hambleton, Ronald K., eds. (1996). Handbook of Modern Item Response Theory. Springer.
- Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., . . . de Vet, H. C. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes.
- O'Malley, A. J., Zaslavsky, A. M., Elliott, M. N., Zaborski, L., & Cleary, P. D. (2005). Case-mix adjustment of the CAHPS Hospital Survey. Health Serv.Res., 40, 2162-2181.
- Snijders, T. A. B., & Bosker, R. J. (1999). Multilevel Analysis: An introduction to basic and advanced multilevel modeling: Sage Publishers.
- Solomon, L. S., Hays, R. D., Zaslavsky, A. M., Ding, L., & Cleary, P. D. (2005). Psychometric properties of a group-level Consumer Assessment of Health Plans Study (CAHPS) instrument. Med Care, 43(1), 53-60.
- Spiegelhalter, D. J. (2005). Funnel plots for comparing institutional performance. Stat Med, 24(8), 1185-1202. doi:10.1002/sim.1970
- Tabachnik, B. G., & Fidell, L. S. (2001). Using Multivariate Statistics (4th ed.): Allyn and Bacon.
- Thompson, B. (2004). Exploratory and confirmatory factor analysis: Understanding concepts and applications. Washington DC: American Psychological Association.
- Zaslavsky, A. M. (2001). Statistical issues in reporting quality data: small samples and casemix variation. Int.J.Qual.Health Care, 13, 481-488.
- Zaslavsky, A. M., Zaborski, L. B., Ding, L., Shaul, J. A., Cioffi, M. J., & Cleary, P. D. (2008). Adjusting Performance Measures to Ensure Equitable Plan Comparisons. Health care financing review, 22, 109-126.

Bijlage I - Keuze voor multi-level modellen als analysemethode

Voor de analyses van patiëntervaringsmetingen wordt bij voorkeur een geavanceerd analysemodel gebruikt: een multi-level regressie model. Dit past namelijk beter bij data over meerdere zorgaanbieders (zorgaanbieders, praktijken, afdelingen) en levert ook betrouwbaarder schattingen per zorgaanbieder op.

Een multi-level model past beter bij de data

Bij vrijwel alle statistische modellen geldt dat deze mogen worden toegepast als aan een aantal voorwaarden (assumpties) is voldaan. Een voorwaarde voor vergelijkende analyses is dat alle waarnemingen / datapunten onafhankelijk moeten zijn. Maar bij het vergelijken van zorgaanbieders gaat het er nu juist om in hoeverre de patiëntervaringen 'afhankelijk' zijn van, of bepaald worden door, de zorgaanbieder of praktijk die de zorg verleend. Het gaat er dus juist om dat de waarnemingen niet onafhankelijk zijn. Enerzijds bieden multi-level modellen een manier om deze afhankelijkheid in kaart te brengen. Anderzijds houden multi-level modellen ook rekening met deze afhankelijkheid bij alle schattingen en resultaten uit het model. Onafhankelijkheid is voor die modellen dan ook geen voorwaarde. Het multi-level model past dus beter bij de data.

De schattingen per zorgaanbieder uit een multi-level model zijn betrouwbaarder

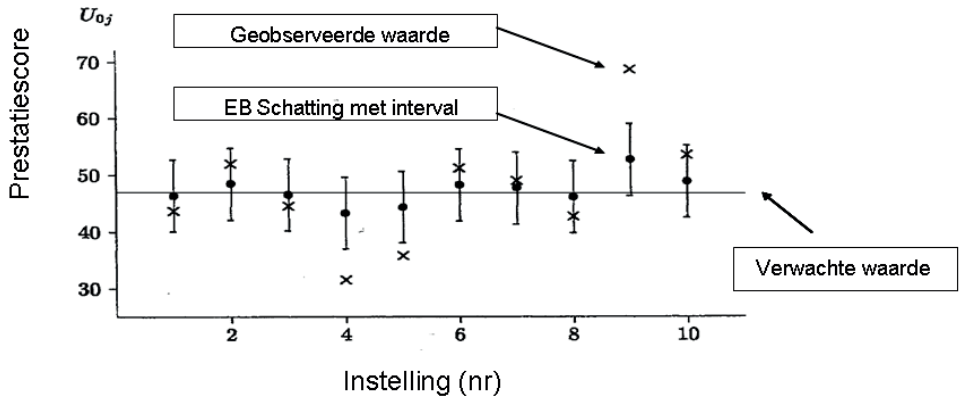
Schattingen per zorgaanbieder uit een multi-level model staan onder verschillende namen bekend: Emperical Bayes schattingen, posterior means of shrinkage estimators. In deze tool spreken we van Emperical Bayes (EB) schattingen. Het gebruik van EB schattingen voor het vergelijken van zorgaanbieders wordt nationaal en internationaal breed gedragen (Arling, Lewis, Kane, Mueller, & Flood, 2007; Damman et al., 2009; De Boer, van der Hoek, Delnoij, & Groenewegen, 2010; Goldstein & Spiegelhalter, 1996; Snijders & Bosker, 1999) Een EB schatting voor een zorgaanbieder is opgebouwd uit twee dingen:

- De geobserveerde waarde. Dit is het gemiddelde over de data van een zorgaanbieder.
- De verwachte waarde. Dit is het gemiddelde over alle zorgaanbieders. Wanneer we niets weten van een zorgaanbieder verwachten we namelijk dat die zorgaanbieder ongeveer gemiddeld zal scoren.

Het gevolg hiervan is dat de EB schatting altijd tussen de verwachte waarde en de geobserveerde waarde ligt (zie ook figuur 1 voor een illustratie; (Snijders & Bosker, 1999)). Belangrijk is natuurlijk de verhouding waarin de geobserveerde waarde en de verwachte waarde vertegenwoordigd zijn in een EB schatting. Voor een belangrijk deel wordt deze verhouding bepaald door het aantal respondenten waarop de geobserveerde waarde voor een zorgaanbieder is gebaseerd.

Wanneer er weinig patiënten/respondenten zijn voor een bepaalde zorgaanbieder kan het gemakkelijk gebeuren dat de geobserveerde waarde voor die zorgaanbieder (deels) op toeval berust. De kans dat toeval een rol speelt is groter naarmate de geobserveerde waarde meer afwijkt van het gemiddelde over alle zorgaanbieders en naarmate deze waarde op minder respondenten is gebaseerd. Kort samengevat: toeval speelt waarschijnlijk een rol bij extreme waarden die gebaseerd zijn op (te) weinig respondenten. Het grote voordeel van de EB schattingen is dat dit een soort correctie vormt voor gevallen waarin het waarschijnlijk is dat toeval een rol speelt. Je zou zelfs kunnen spreken van een correctie voor te kleine aantallen. Zaligmakend zijn de EB schattingen nu ook weer niet. Een nadeel is dat kleine zorgaanbieders die niet kunnen voldoen aan de vereiste steekproefgrootte vaak een score dicht bij het gemiddelde krijgen (De Boer et al., 2010). Als zij werkelijk heel slecht of heel goed zouden presteren, wordt dit voor een deel weggepoetst door de EB schattingen.

Figuur 1: Een illustratie van Emperical Bayes (EB) schattingen.



-

i PREMs: Patient Reported Experience Measures zijn de door patiënten/cliënten gerapporteerde ervaringen met het zorgproces

ii PROMs: Patient Reported Outcome Measures zijn de door patiënten/cliënten gerapporteerde uitkomsten van de zorg

iii Imputatie: Het bepalen en introduceren van een nieuwe waarde op een plaats waar een waarde ontbreekt of mist

iv Factoranalyse: een statistische techniek waarmee voor een groot aantal vragen een beperkter aantal achterliggende variabelen (factoren) kan worden geïdentificeerd.

v Betrouwbaarheidsanalyse: een analyse die inzicht geeft in de mate waarin (groepen) vragen vrij zijn van meetfout

vi Reguliere meting: Een meting met een reeds ontwikkelde vragenlijst die niet primair tot doel heeft de vragenlijst verder te ontwikkelen

vii Thema: Een groep vragen die inhoudelijk over eenzelfde onderwerp gaan en waarover een gemiddelde kan worden berekend

viii Case-mix adjusters: patiëntkenmerken die vergelijkingen tussen zorgaanbieders vertekenen en waarvoor een statistische correctie wordt toegepast bij het vergelijken van zorgaanbieders.

ix Intraclass correlatie coefficient (ICC): een getal tussen 0 en 1 dat aangeeft in welke mate individuen binnen een groep (zorgaanbieders) meer op elkaar lijken dan individuen tussen groepen. Bij het vergelijken van patiëntervaringen tussen instellingen geldt dat naarmate de ICC hoger is, de verschillen tussen instellingen groter zijn.

x Reliability: Een maat voor de betrouwbaarheid van de van de geschatte score van een groep (zorgaanbieder) op basis van een multi-level model.