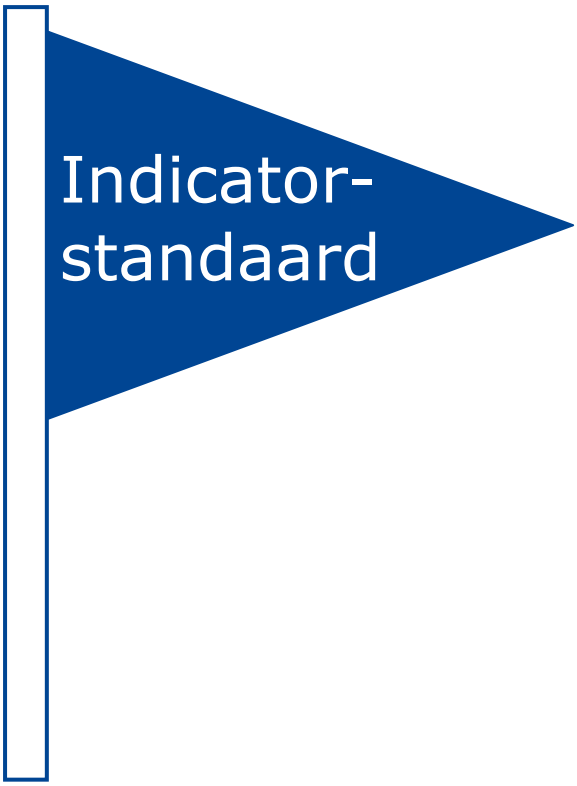
VU Research Portal

Indicatorstandaard 2.0

Koolman, Xander; Zuidgeest, M; Visser, Johan; Appelman, M.

2012

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



Methodologische criteria voor de ontwikkeling van betrouwbare kwaliteitsindicatoren in de zorg

versie 2.0
31 december 2012

Colofon

Opdracht en verantwoordelijkheid
Tot stand gekomen in opdracht
van het Kwaliteitsinstituut i.o. / CVZ

Auteurs versie 2.0
dr. Xander Koolman (SiRM)
dr. Marloes Zuidgeest (SiRM)
drs. Johan Visser (SiRM)
drs. Marja Appelman (SiRM)

Auteurs versie 1.5
dr. Xander Koolman (SiRM)
drs. Johan Visser (SiRM)
drs. Marja Appelman (SiRM)

Auteurs versie 1.0
drs. Nicoline Beersen (KPMG-Plexus)
dr. Marc Berg (KPMG-Plexus)
dr. Xander Koolman (SiRM)

Inhoudsopgave

1. Doel en kader indicatorstandaard

De overheid is verantwoordelijk voor de transparantie in de zorg. Dit volgt uit haar verantwoordelijkheid voor het goed functioneren van het zorgstelsel. Zonder voldoende informatie over de kwaliteit van zorg kan de patiënt of cliënt geen goede keuze maken en worden zorgverzekeraars niet gestimuleerd tot het inkopen van goede zorg. Keuze-informatie over de kwaliteit van zorg – onder meer via indicatoren - vormt een van de speerpunten van het Kwaliteitsinstituut. Transparantie is een voorwaarde voor een goed functionerend zorgstelsel.

De concrete invulling en het genereren van kwaliteitsinformatie vertrouwt de overheid aan partijen in de zorgtoe. Om tot betrouwbare en vergelijkbare kwaliteitsinformatie te komen dienen onder meer methodologische criteria te worden afgewogen. De methodologische criteria voor indicatoren over de kwaliteit van zorg zijn uitgewerkt in deze indicatorstandaard. De indicatorstandaard is geschreven voor experts die de indicatoren ontwikkelen, of deze ontwikkeling begeleiden. De indicatorstandaard sluit aan bij actuele wetenschappelijke inzichten.

De indicatorstandaard wordt vergezeld door de routekaart. Deze kaart neemt partijen in de zorg mee in afwegingen en keuzes die gemaakt dienen te worden tijdens de ontwikkeling en verbetering van indicatoren. Daarbij is aandacht voor de potentiële impact van de keuzen op de kwaliteit van de indicator(waarde).

1.1. Doel indicatorstandaard

Het doel van de indicatorstandaard is een toetsingskader voor kwaliteitsindicatoren te bieden en partijen in de zorg ondersteuning te bieden bij:

1. (door)ontwikkeling van indicatoren
2. toetsing van indicatoren

De methodologische criteria in de indicatorstandaard zijn gelijk voor zowel zorginhoudelijke indicatoren als patiëntervaringen, ook wel klantervaringsindicatoren genoemd. De criteria gelden voor alle sectoren, maar enkele zijn niet voor elk type indicator relevant. De standaard beschrijft tevens hoe de gebruiker van kwaliteitsinformatie dient te worden geïnformeerd over de zeggingskracht van de indicatorwaarden.

1.2. Kaders en positie van de indicatorstandaard

De wetgever heeft in verschillende wetten[1] als norm geformuleerd dat de zorgaanbieder "betrouwbare en vergelijkbare kwaliteitsinformatie" openbaar dient te maken. Deze indicatorstandaard beschrijft de belangrijkste toetsingscriteria voor betrouwbare en vergelijkbare kwaliteitsinformatie vanuit een wetenschappelijke invalshoek. De indicatorstandaard is onderdeel van het toetsingskader van het Kwaliteitsinstituut i.o.

De indicatorstandaard bouwt voort op de Indicatorstandaard 1.5 en de daar aan voorafgaande Indicatorstandaard 1.0, het Aire instrument[2] en de handboeken van het Centrum Klantervaring Zorg (CKZ)[3]. Deze onderliggende documenten bevatten aanvullende adviezen voor het ontwikkelen en bewerken van indicatoren.

[1] Bepalingen zoals opgenomen in de Wet Marktwerking Gezondheidszorg en de Wet toelating zorginstellingen.
[2] Appraisal of Indicators through Research and Evaluation (AIRE); versie 2.0, 2007, Johan de Koning, Annke Smulder, Niek Klazinga, Afdeling Sociale Geneeskunde van het Academisch Medisch Centrum van de Universiteit van Amsterdam
[3] "CQI Ontwikkeling: richtlijnen en voorschriften voor de ontwikkeling van een CQI meetinstrumen" 2008 en "Eisen & Werkwijzen CQI Metingen", 2011, beide van het Centrum Klantervaring Zorg, Utrecht.

1.3. Gebruikersdoel: keuze-informatie

De ontwikkeling en toetsing van kwaliteitsindicatoren is afhankelijk van het gebruiksdoel. Kwaliteitsindicatoren ten behoeve van het toezicht door de Inspectie van de Gezondheidszorg (IGZ) worden per definitie door de IGZ bepaald. Zorgaanbieders en zorgverzekeraars worden in staat geacht om – zonder of met geringe overheidsbemoeienis – hun informatievraag te definiëren en beantwoord te krijgen. Bij zorggebruikers ligt het anders. De overheid zal daarom bij het evalueren van resultaten en mogelijke ingrepen in het transparantieprogramma nadrukkelijk aandacht besteden aan de totstandkoming van voldoende keuze-informatie voor zorggebruikers[4]. Zorggebruikers zijn te onderscheiden in kiezende toekomstige zorggebruikers, huidige zorggebruikers, vertegenwoordigers van zorggebruikers en verwijzers.

In deze indicatorstandaard gebruiken we de term 'patiënt' wanneer we spreken over zorggebruikers, zoals gebruikelijk is in het grootst deel van de curatieve zorg. In delen van de zorg, zoals de AWBZ, is het gebruik om te spreken over 'cliënt'. Overal in het rapport waar 'patiënt' wordt ook 'cliënt' bedoeld.

Bijlage 3 bevat informatie voor het gebruik van indicatoren door overige gebruikersdoelen.

1.4. Typen indicatoren: proces, structuur en uitkomst

In deze indicatorstandaard wordt onderscheid gemaakt tussen structuur-, proces- en uitkomstindicatoren.

- Structuurindicatoren betreffen vragen naar de organisatorische randvoorwaarden van de zorg, waardonder de aanwezigheid van registraties en kwalificaties van personeel.
- Procesindicatoren beschrijven metingen aan de procesgang van de zorg, waaronder wacht- en doorlooptijden, aanrijdtijden, volume-indicatoren, risico-evaluaties, indicatiestelling, praktijkvariatie en richtlijnconformiteit.
- Uitkomstindicatoren zijn metingen van uitkomsten van geleverde zorg (zoals patiënttevredenheid, 10-jaars overleving, mate van ziekteactiviteit, kwaliteit van leven, complicaties, sterfte).

Uitkomstindicatoren hebben de grootste informatieve waarde. Ze geven tenslotte een indicatie van de daadwerkelijke resultaten van de geleverde zorg. Daarbij wordt onderscheid gemaakt tussen zorginhoudelijke indicatoren die samenhangen met de gezondheidsuitkomst en patientenervaringen. Patientenervaringen zijn uitkomstindicatoren omdat zij een resultaat van de zorg beschrijven dat direct door de patient wordt gewaardeerd ook al blijft de gezondheidsuitkomst gelijk. Hiermee verschillen patientervaringen van bijvoorbeeld procesindicatoren die door de patient worden gewaardeerd vanwege hun relatie met goede gezondheidsuitkomsten.

1.5. Proces van indicatorstandaard 1.0 naar versie 2.0

Als onderdeel van de ontwikkeling van de Indicatorstandaard 1.5 is de eerste indicatorstandaard geëvalueerd. Daarbij is een selectie van stuurgroeppartijen geïnterviewd. Bij de selectie is gestreefd naar een goede vertegenwoordiging van gebruikers (patiëntenverenigingen en zorginkopers) en zorgaanbieders. Versie 2.0 bevat een volwaardig hoofdstuk over steekproef en responsvergelijkbaarheid en is verder aangepast zodat patiëntervaringen nu volledig geintegreerd zijn in de indicatorstandaard. Daarnaast is de tekst aangepast aan de veranderde beleidscontext waarin de standaard zal worden gebruikt.

Daarnaast zijn de Statistiek Commissie van het CKZ, de commissie van deskundigen Transparantie van de Zorg en de Stuurgroep Transparantie Zorg geconsulteerd. Op basis van de inbreng van deze groepen is de indicatorstandaard in twee stappen geactualiseerd. De indicatorstandaard is daarmee onderdeel van een lerend systeem waaraan onderzoekers, ontwikkelaars, gebruikers, en

[4] Inhoudelijke kaderstelling voor het transparantieprogramma (tweede concept), 28 november 2011, Ministerie van VWS.

zorgaanbieders deelnemen. Voor een meer gedetailleerde beschrijving van de betrokkenen en de evaluatie en deelnemende partijen zie bijlagen 1 en 2.

1.6. De beoordelingscriteria

In de internationale literatuur bestaat geen gouden standaard voor het ontwikkelen, meten en openbaar maken van kwaliteitsindicatoren[5]. Wel is veel gepubliceerd over het concept kwaliteit, het meten van kwaliteit en de afzonderlijke kwaliteitscriteria waaraan goede indicatoren dienen te voldoen. De indicatorstandaard bouwt voort op deze uitgebreide literatuur.

Kwaliteitsindicatoren en vragenlijsten worden ontwikkeld door epidemiologen, artsen, psychologen, gezondheidzorgonderzoekers, economen en anderen. Elke groep heeft een eigen begrippenkader. De concepten sluiten in veel gevallen aan, maar zijn soms ook wezenlijk verschillend. In deze indicatorstandaard wordt een begrippenkader geïntroduceerd dat nauw aansluit bij dat van de epidemiologen[6], omdat deze groep een groot deel van de indicatoren heeft geleverd. Het economische begrippenkader sluit nauw aan bij het epidemiologische kader. Het psychologische kader wijkt echter af, vooral omdat het onderscheid tussen betrouwbaarheid en validiteit minder strikt is dan in de andere wetenschapsgebieden.

In het epidemiologische begrippenkader bestaan drie vormen van vertekening (bias): *measurement error/information bias, confounding bias* en *selection bias*. Toegepast op kwaliteitsmeting in de gezondheidszorg vanuit het patiëntperspectief kunnen we deze begrippen vertalen naar registratievergelijkbaarheid, populatievergelijkbaarheid en steekproefvergelijkbaarheid. Indien de indicatorwaarden vertekend zijn, dan is de vergelijkbaarheid van de indicatorwaarden verminderd.

Een vergelijkbare indicator is pas bruikbaar indien deze een relatie heeft met kwaliteit van zorg en slechts beperkt gevoelig is voor toeval. Daarom gelden de volgende beoordelingscriteria:

1. inhoudsvaliditeit;
2. vertekening;
    a. registratievergelijkbaarheid (*measurement bias*);
    b. populatievergelijkbaarheid (*confounding bias*);
    c. steekproef- en responsvergelijkbaarheid (*selection bias*);
3. statistisch betrouwbaar onderscheiden.

Hieronder worden deze termen kort toegelicht.[7]

1. **Inhoudsvaliditeit**

   Er bestaat een duidelijke relatie tussen de geleverde zorg (of het ontbreken daarvan) en de zorguitkomsten. Voor *uitkomstindicatoren* betekent dit dat de gemeten uitkomst aantoonbaar beïnvloedbaar is door de zorgaanbieder(s) waar de indicator betrekking op heeft. Voor *structuur- en procesindicatoren* betekent dit dat is aangetoond dat de gemeten structuur of processen de gewenste zorguitkomsten beïnvloeden. Inhoudsvaliditeit wordt zowel beoordeeld op het niveau van de indicator als op het niveau van de indicatorset. De inhoudsvaliditeit op setniveau is goed als de set van indicatoren de relevante domeinen van de geleverde zorg goed dekt. Afhankelijk van de visie op kwaliteit wordt tevens gelet op de relevante fasen (preventie, indicatie, proces van zorg zelf, uitkomsten) en de kwaliteitsdomeinen (effectiviteit, veiligheid, patiëntgerichtheid).

2. **Vertekening**

[5] Het Centrum voor Klantervaring in de Zorg (CKZ) beheert en ontwikkelt een standaard voor het ontwikkelen, meten en openbaar maken van vragenlijsten voor het meten van klantervaringen. Alle vragenlijsten en meetresultaten die aan de CQI-standaard voldoen ontvangen van het CKZ een keurmerk.
[6] Zie hoofdstuk 12 uit Modern Epidemiology, Kenneth J. Rothman, Sander Greenland, Timothy L. Lash, Lippincott Williams & Wilkins; Third edition, 2008
[7] Zie bijlage 3 voor een verklarende woordenlijst.

2a. **Registratievergelijkbaarheid**
De vergelijkbaarheid van indicatorwaarden wordt niet beïnvloed door verschillen tussen de geregistreerde en de werkelijke waarden van structuur, proces, uitkomst en populatie. Andere termen die worden gebruikt zijn meetvertekening, classificatievertekening, testvertekening, respons heterogeniteit en differential item functioning.

2b. **Populatievergelijkbaarheid**
De vergelijkbaarheid van indicatorwaarden wordt niet beïnvloed door de verschillen in populatiekenmerken. De berekende indicatorwaarden weerspiegelen daadwerkelijke verschillen in de kwaliteit van de geleverde zorg en niet de verschillen in de populatiekenmerken van de zorgaanbieders. De correctie voor verschillen in populatiekermerken wordt 'casemix-correctie' of 'risk adjustment' genoemd.

2c. **Steekproef- en responsvergelijkbaarheid**
De vergelijkbaarheid van indicatorwaarden wordt niet beïnvloed door verschillen tussen de steekproef en de totale behandelde populatie. Met andere woorden: de steekproef is representatief voor de gehele behandelpopulatie waarop de indicator van toepassing is.

3. **Statistisch betrouwbaar onderscheiden**
Een indicator dient het vermogen te hebben om zorgaanbieders met bovengemiddelde en ondergemiddelde indicatorwaarden te onderscheiden van gemiddeld scorende aanbieders. Bij dit criterium speelt het aantal patiënten waarover een zorgaanbieder gegevens kan aanleveren een essentiële rol: bij een te laag aantal waarnemingen wordt de rol van toeval veelal te groot om betrouwbare verschillen in indicatorwaarden te kunnen onderscheiden.

Een ideale kwaliteitsindicator is inhoudsvalide, niet vertekend en weinig gevoelig voor toeval. De criteria zijn afzonderlijk te beoordelen maar beïnvloeden elkaar. Tijdens de beoordeling van een criterium wordt echter verondersteld dat de indicator(set) goed scoort op de overige criteria. De beoordelingscriteria zijn algemeen toepasbaar en bruikbaar voor de toetsing van afzonderlijke zorginhoudelijke indicatoren, alsmede voor de toetsing van indicatoren die worden gevuld via vragenlijsten zoals de Consumer Quality Index (CQI) en de Patiënt gerapporteerde uitkomst maten (PROM).

1.7. Het gaat om de combinatie van de beoordelingscriteria

Essentieel bij de toepassing van de indicatorstandaard is het inzicht dat voor wat betreft de afzonderlijke criteria *de zwakste schakel de uiteindelijke kwaliteit van de indicator bepaalt*. De inhoudsvaliditeit van een indicator kan goed zijn, maar als de registratievergelijkbaarheid slechts deels voldoet, dan is de gebruikswaarde van de indicator gering. Zijn er meerdere zwakke plekken (bijvoorbeeld naast een 'voldoet deels' oordeel op registratievergelijkbaarheid ook nog een 'voldoet deels' oordeel op vergelijkbaarheid), dan *is de uiteindelijke kwaliteit zwakker dan de zwakste schakel.*

Het is mogelijk dat een indicator die afzonderlijk 'voldoet deels' scoort op enkele criteria, in gezamenlijkheid met andere indicatoren beter scoort. De indicatorstandaard richt zich echter op de beoordeling van de afzonderlijke indicator. Dat geldt ook voor de oordelen in de vorm van signaalvlaggen groen, geel en rood. Als gevolg kan een indicator in combinatie met andere indicatoren bruikbaar zijn, terwijl de afzonderlijke indicator dat niet is.

Een overzicht van een indicatorset-beoordeling wordt gegeven in tabel 1. In het overzicht staan zeven verschillende signaalvlaggen: (1) inhoudsvaliditeit op setniveau, (2) inhoudsvaliditeit op indicatorniveau, (3) populatievergelijkbaarheid op indicatorniveau, registratievergelijkbaarheid op (4) indicator en op (5) indicator per zorgaanbiederlokatie of informatie-eenheid (IE) en (6)

statistisch betrouwbaar onderscheiden op indicator en op (7) indicator per informatie eenheid (IE) niveau.

De relevantie van de criteria is afhankelijk van het type indicator. Een structuurindicator is niet gevoelig voor populatiekenmerken en toeval en wordt daarom niet voorzien van een signaalvlag voor deze criteria. In tabel 1 worden vier typen indicatoren onderscheiden: structuurindicatoren (a), procesindicatorren die niet gevoelig zijn voor populatiekenmerken (b), procesindicatoren die wel gevoelig zijn voor populatiekenmerken (c), en uitkomstindicatoren (d).

	Indicator-set	Indicator	Indicator per IE*
Inhoudsvaliditeit	a, b, c, d	a, b, c, d	
Populatievergelijkbaarheid		c, d	
Registratievergelijkbaarheid		a, b, c, d	b, c, d
Steekproef- en responsvergelijkbaarheid		c, d	
Statistisch betrouwbaar onderscheiden		c, d	c, d

*IE, Informatie Eenheid; a=structuurindicatoren, b=procesindicatoren die niet gevoelig zijn voor populatiekenmerken, c= procesindicatoren die wel gevoelig zijn voor populatiekenmerken, en d=uitkomstindicatoren

Tabel 1: Overzicht signaalvlaggen

De beoordelingscriteria van de indicatorstandaard kunnen op verschillende momenten op indicatoren worden toegepast. Het tijdstip heeft geen invloed op de keuze van criteria die worden gehanteerd, maar wel op de informatie die beschikbaar is om elk criterium te toetsen.

Een indicator kan beoordeeld worden tijdens de ontwikkeling van de indicator(set) en zonder dat er daadwerkelijk data verzameld zijn. Dit noemen we de ex-ante beoordeling van de indicator(set). De beoordeling kan ook plaatsvinden na de data-uitvraag of een praktijktest. De resultaten van de meting kunnen dan bij de beoordeling van de kwaliteit van de indicator worden betrokken. Dit noemen we een ex-post beoordeling van de indicator(set).

1.8. Expertise nodig voor toepassen beoordelingscriteria

Of een indicator voldoet aan de criteria kan getoetst worden met voorwaarden. Deze voorwaarden vormen geen scorelijstje dat een leek in staat stelt een indicator(set) te beoordelen. Slechts over het statistisch betrouwbaar onderscheiden kunnen oordelen worden toegekend op basis van eenvoudige beslisregels, maar bij de andere criteria is het oordeel van experts nodig.

De vereiste kennis van deze experts verschilt tussen de criteria. Een oordeel over de inhoudsvaliditeit en populatievergelijkbaarheid vereist een combinatie van medisch inhoudelijke, wetenschappelijke en zorginhoudelijke experts. Een oordeel over de registratie en steekproef- en responsvergelijkbaarheid vereist vooral kennis van de dataverzameling en algemene methodische kennis.

In afwezigheid van wetenschappelijke literatuur over de relatie tussen proces en uitkomst, bijvoorbeeld, kan niet worden gesteld dat een procesindicator ‘niet inhoudelijk valide’ is. In sommige gevallen is er wel consensus tussen experts dat de relatie er is, maar is er eenvoudigweg (nog) geen onderzoek gedaan. Ook kan de relatie als zodanig evident worden gezien door experts dat er geen onderzoek nodig is. In dat geval is de inhoudsvaliditeit voor verbetering vatbaar, maar

de beoordeling of de inhoudsvaliditeit dan toch ‘genoeg’ is of echt ‘te weinig’ is daarmee altijd een expertoordeel.

2. Inhoudsvaliditeit

2.1. Definities

Inhoudsvaliditeit op	
Setniveau	De indicatorset geeft een duidelijke relatie weer tussen de geleverde zorg en de patiëntervaringen en zorguitkomsten.
structuur- of procesindicator	Het is aangetoond dat de gemeten structuur of processen daadwerkelijk de gewenste zorguitkomsten beïnvloeden.
uitkomstindicator	De gemeten uitkomst heeft een duidelijke relatie met de geleverde zorg.

Inhoudsvaliditeit heeft betrekking op de indicatorset (zorginhoudelijke en patiëntervaringsindicatoren) als geheel *en* op de afzonderlijke indicatoren in een set. Inhoudsvaliditeit vereist een relatie tussen de kwaliteit van de geleverde zorg (of het ontbreken daarvan) en de zorguitkomsten. De inhoudsvaliditeit van een indicator of set indicatoren refereert aan de mate waarin de indicator(set) daadwerkelijk de kwaliteit van zorg meet. Dit heet ook wel de *content validity* van de indicator.

Bij de beoordeling van de inhoudsvaliditeit is het meestal afdoende om stil te staan bij de vraag *of* er wetenschappelijke literatuur bestaat waarin de relatie van de indicator tot een zorguitkomst wordt aangetoond. In sommige gevallen (daar waar twijfels bestaan over de kwaliteit of relevantie van deze literatuur, bijvoorbeeld) kan dieper worden gegraven naar de wijze *waarop* deze indicatoren zijn of (moeten) worden gevalideerd. Dan is er sprake van *construct* en/of *criteriumvaliditeit*. Informatie hierover staat in bijlage 5.

In onderstaande paragrafen staat de inhoudsvaliditeit uitgewerkt voor de indicatorset als geheel en de afzonderlijke indicatoren.

2.2. Omschrijving van goede kwaliteit van zorg als basis voor inhoudsvaliditeit

Om te komen tot een inhoudsvalide set indicatoren dienen vartijen in de zorg te beschikken over één visie op kwaliteit van zorg: een omschrijving wat goede zorg is. Het uitgangspunt van de indicatorstandaard is dat partijen in de zorg beschikken over een omschrijving van goede zorg, in bijvoorbeeld een kwaliteitsvisie of professionele standaard, die uitgaat van het patiëntkeuzeperspectief. Voor gebruikers van indicatorwaarden met een ander doel is in bijlage 3 informatie opgenomen waarmee zij de indicatoren optimaal kunnen gebruiken.

Een goede omschrijving van de kwaliteit van zorgis een voorwaarde voor bruikbare indicatoren. Indien de omschrijving niet overeenkomt met kwaliteit van zorg vanuit het perspectief van de patiënt, dan kan een indicator(set) goed scoren op alle criteria van de indicatorstandaard en toch weinig gebruikswaarde hebben.

De omschrijving van de kwaliteit van zorg zal waarschijnlijk per zorgdomein en fase in het zorgproces de meest relevante uitkomsten en patiëntervaringen bevatten. De omschrijving dient ontwikkeld te worden met input van patiënten. Daarbij kan een patiëntenorganisaties worden ondersteund met wetenschappelijk voorwerk bij het formuleren van een visie op kwaliteit. Hierdoor kunnen de kwaliteitsvisies tegemoet komen aan de specifieke verschillen tussen sectoren, maar blijven de overwegingen en terminologie blijven op elkaar aansluiten.

2.3. Inhoudsvaliditeit in relatie tot het type indicator

Voor alle typen indicatoren geldt dat de inhoudsvaliditeit essentieel is. Dit betekent voor *structuur-* en *procesindicatoren* dat moet zijn onderbouwd dat de aanwezigheid van een bepaalde structuur of het uitvoeren van een bepaald proces de zorguitkomsten beïnvloedt.

De relatie tussen structuur- en procesindicatoren en zorguitkomsten verdient zorgvuldige onderbouwing. In veel gevallen is een sterke relatie tussen de processen en uitkomsten aangetoond in een onderzoekscontext. Indien de rapportage van de processen deel uitmaakt van een verplichte verantwoording met consequenties voor de financiën of reputatie, dan blijkt de relatie tussen processen en uitkomsten echter zeer zwak of afwezig (Werner 2006; Nicholas 2010). Het verdwijnen van de relatie tussen procesindicatoren en uitkomsten kan samenhangen met het gedrag dat openbare publicatie oproept. Zo wordt voor een deel van de zorgaanbieders niet de uitkomst maar het goed scoren op de procesindicator het doel. Dit kan leiden tot pervers gedrag.

Een voorbeeld is wanneer de aanrijdtijd van een ambulance wordt gemeten vanaf de melding, onder voorwaarde dat een ambulance aanwezig is op het terrein. Onder die voorwaarde leidt het stallen van ambulances buiten het terrein tot een verbetering van de procesindicatorwaarde, maar een verslechtering van de geboden zorg.

Uitkomstindicatoren zijn inhoudsvalide wanneer is onderbouwd dat het handelen van de zorgaanbieders een relatie heeft met de uitkomstindicatorwaarde of patiëntervaring. De relatie tussen de uitkomstindicator en de uitkomsten van zorg is veelal direct duidelijk. Mede daarom is er onder de gebruikers van de resultaten meer intrinsieke interesse voor uitkomstindicatoren (Mant 2001). Een uitkomstindicator kan echter gevoelig zijn voor vertekening en toeval (zie hoofdstuk 3 tot en met 6). Bij de beoordeling van de inhoudsvaliditeit wordt echter verondersteld dat dit niet het geval is.

2.4. Beoordelen inhoudsvaliditeit op setniveau

Bij de inhoudsvaliditeit op setniveau wordt een inschatting gemaakt over de 'volledigheid' van de set: kan met de set als geheel een indruk worden verkregen van de voor patiënten relevante kwaliteit van zorg, zowel zorginhoudelijke kwaliteit als patiëntervaringen. Tijdens de toetsing wordt verondersteld dat de afzonderlijke indicatoren inhoudsvalide en niet vertekend zijn (hoofdstuk 3 tot en met 5).

De mate waarin een set voldoende dekkend is hangt af van de visie op kwaliteit die dient te worden gedekt. Indien een visie op kwaliteit bestaat uit fasen van zorg (bijvoorbeeld preventie, indicatiestelling, proces en uitkomst) en domeinen per fase, dan kan met behulp van een domein/fase-matrix worden onderzocht in welke mate een set dekkend is.

Bij het selecteren van indicatoren die moeten worden ontwikkeld en geïmplementeerd dient een inschatting te worden gemaakt van de informatieve waarde voor zorggebruikers en administratieve lasten voor zorgaanbieders[8]. In de routekaart bij de indicatorstandaard staan afwegingen hiervoor beschreven. Indien de administratieve lasten aanleiding geven om de set niet verder uit te breiden, dan beïnvloedt dat het oordeel over de inhoudsvaliditeit echter niet.

Een deel van de indicatoren is niet afzonderlijk te beoordelen. Zo kan een indicator gericht op valincidenten in een verpleeghuis samenhangen met het fixatiebeleid. Een toename van fixatie kan een ongewenst effect zijn van het streven naar minder valincidenten. In dergelijke gevallen is het aan te raden om beide indicatoren deel te laten uitmaken van de indicatorset.

[8] Inhoudelijke kaderstelling voor het transparantieprogramma (tweede concept), 29 november 2011, Ministerie van VWS

Vlag	Toelichting	Voorbeeld
Voldoet	Alle relevante zorgonderdelen zijn opgenomen in de indicatorset de uitkomsten die relevant zijn.	De indicatorset longcarcinoom bevat indicatoren over de diagnostiek, het behandelproces en de behandeluitkomsten. Zorginhoudelijk is de inhoudsvaliditeit goed. De patiëntervaringen worden gemeten met bijvoorbeeld de CQ index methodiek.
Voldoet deels	Enkele relevante kwaliteitsdomeinen zijn in de set opgenomen, maar belangrijke aspecten ontbreken. De patiëntervaringen worden uitgevraagd.	De huidige indicatorset Ziekten Adenoid en tonsillen richt zich op relevante kwaliteitsaspecten: vooral het perioperatieve proces (incl. enkele uitkomstmaten). Uitkomstindicatoren voor diagnostiek en behandeling ontbreken. Patiëntervaringen worden uitgevraagd.
Voldoet niet	De indicatorset omvat geen of nauwelijks relevante kwaliteitsdomeinen en patiëntervaringen worden uitgevraagd. Of De indicatorset omvat enkele relevante kwaliteitsdomeinen, maar patiëntervaringen worden niet uitgevraagd.	De set chronische rhinosinusitis bestaat uit één proces en één structuurindicator, beiden gericht op de beeldvormende diagnostiek. Indicatoren op het gebied van uitkomst en indicatoren op het gebied van indicatiestelling ontbreken. Patiëntervaringen worden niet uitgevraagd.

Tabel 2 Signaalvlaggen inhoudsvaliditeit op setniveau

2.5. Beoordelen inhoudsvaliditeit op indicatorniveau

De inhoudsvaliditeit van een indicator is goed als er een duidelijke relatie is tussen de geleverde zorg en de zorguitkomsten. Voor uitkomstindicatoren betekent dit dat is aangetoond dat de gemeten uitkomst of patiëntervaring beïnvloedbaar is door de zorgaanbieder(s) waar de indicator betrekking op heeft. Voor structuur- en procesindicatoren betekent dit dat is aangetoond dat de gemeten structuur of processen ook daadwerkelijk de gewenste zorguitkomsten beïnvloeden.

De achterliggende gedachte is dat de zorgaanbieders kwaliteit leveren die de indicatorwaarde bepaalt. Hoe meer de kwaliteit van zorg de indicatorwaarde bepaalt, hoe hoger de inhoudsvaliditeit. Bij de beoordeling van de inhoudsvaliditeit dient te worden verondersteld dat de registratie, de populatie en de steekproef vergelijkbaar zijn (gemaakt) en dat toeval geen rol speelt (zie hoofdstuk 3 t/m 6).

Net zoals bij het opstellen van richtlijnen voor medisch handelen kunnen voor de zorginhoudelijke indicatoren 'gradaties van bewijskracht' (hierarchy of evidence) van wetenschappelijke studies worden gehanteerd bij het onderbouwen van de relatie tussen zorgproces en zorguitkomst. Hierbij moet wel in acht worden genomen dat er voor *handelingen* (zoals omschreven in de richtlijnen)

meer onderzoek voorradig is dan voor het onderbouwen van de validiteit van een indicator. Voor de gradaties geldt 'hoe hoger hoe beter' (CBO 2005):

A1 gebaseerd op systematische reviews die in ten minste enkele onderzoeken van A2-niveau betreffen, waarbij de resultaten van de afzonderlijke onderzoeken consistent zijn;

A2 gebaseerd op gerandomiseerd vergelijkend klinisch onderzoek van goede kwaliteit van voldoende omvang en consistentie;

B gebaseerd op gerandomiseerde klinische trials van matige kwaliteit of onvoldoende omvang of ander vergelijkend onderzoek (=niet-gerandomiseerd, vergelijkend cohort-onderzoek, patiëntcontrole-onderzoek);

C gebaseerd op niet-vergelijkend onderzoek;

D consensus van deskundigen.

Indien wetenschappelijk bewijs voor de relatie tussen zorgproces en –uitkomst niet eenduidig is, of als er geen onderzoek naar is gedaan, kan een groep experts, samengesteld op basis van (wetenschappelijke) ervaring, worden geraadpleegd om consensus over afzonderlijke indicatoren en/ of een indicatorset als geheel te bereiken. Die expertgroep dient te beschikken over: medisch-wetenschappelijke kennis, zorginhoudelijke kennis, statistische kennis, kennis van kwaliteitsmeting en gebruik van indicatoren.

Daarbij kan gebruik gemaakt worden van een consensusmethodiek. Met een consensus methodiek worden de verschillende individuele opinies en ervaringen verkend (consensus measurement) en gebruikt om tot een groepsopinie te komen (consensus development) (Jones 1995, Fink 1984). Een groepsbeoordeling of opinie heeft de voorkeur boven beoordelingen of opinies van individuele experts omdat deze consistenter zijn en minder gevoelig voor persoonlijke bias en gebrek aan reproduceerbaarheid (Campbell *et al.* 2002). Indien de methoden hierboven niet volstaan kan worden gekozen voor het opzetten van een validatiestudie.

In de onderstaande tabel staat beschreven en met voorbeelden geïllustreerd hoe het criterium inhoudsvaliditeit op indicatorniveau beoordeeld kan worden.

Vlag	Toelichting	Voorbeeld
Voldoet	Een eenduidige relatie tussen de zorginhoudelijke uitkomstindicatoren en het zorgproces is beschreven. Of, in het geval van een proces- of structuurindicator, de relatie tussen het structuur- of proces-kenmerk en de zorguitkomsten is beschreven. Deze relatie is onderbouwd door wetenschap-pelijk onderzoek (ten minste graad C bewijskracht), of, als dit onderzoek niet voorhanden is, op basis van consensus (graad D bewijskracht) onder deskundigen middels een voorgeschreven methodiek. Voor patiëntervaringsindicatoren geldt dat deze zijn geselecteerd in samenspraak met belanghebbende partijen en dat een belangscore aangeeft dat	Voor de ziekenhuis indicator 'Percentage patiënten met een herseninfarct dat binnen 1 uur na binnenkomst in het ziekenhuis is behandeld met trombolyse' (zgn. 'door-to-needle' time; indicatorset Beroerte) is de relatie tussen 'procesmaat' (de door-to-needle time) en 'zorguitkomst' (uiteindelijke overleving en functionele status) duidelijk gelegd. De relatie wordt uitgebreid met literatuur referenties onderbouwd.

Vlag	Toelichting	Voorbeeld
	de indicator een wezenlijk belang dient (graad D bewijskracht).	
Voldoet deels	Een eenduidige relatie tussen de uitkomstindicator en het zorgproces is niet beschreven. Of, in het geval van een proces- of structuurindicator, de relatie tussen het structuur- of proces-kenmerk en de zorguitkomsten is niet beschreven. Tegelijkertijd is er geen reden om aan het bestaan van deze relatie te twijfelen. *Of:* De relatie is onderbouwd door wetenschappelijk onderzoek (ten minste graad C bewijs-kracht), en er is een matige of een niet eenduidige relatie tussen het zorgproces en de zorguitkomsten. *Of:* De relatie is onderbouwd door consensusoordeel van deskundigen, de gevolgde consensus methodiek is echter onvoldoende.	Veel indicatoren over patiëntgerichtheid vallen in deze categorie. Indicatoren over 'doorlooptijd' en 'wachttijden', bijvoorbeeld, gaan er vanuit dat kortere doorloop- en wachttijden te prefereren zijn. Dit is in de meest gevallen ook zo, maar onderbouwing van dergelijke aannames is altijd te prefereren. Sommige indicatoren vragen naar 'het aanwezig zijn van een lokaal protocol' bij een zorgaanbieder. Het is echter uit de literatuur bekend dat het 'hebben' van een protocol (de structuur-indicator) niet betekent dat er daadwerkelijk wordt 'gehandeld conform het protocol'. De consensus kan onvoldoende zijn als het panel van experts geen goede afspiegeling is van de zorgprofessionals die betrokken zijn bij de behandeling van een bepaalde aandoening. Voor patiëntervaringsindicatoren zijn niet alle belanghebbenden betrokken bij de selectie, of zijn de belangscores niet uitgevraagd.
Voldoet niet	Een eenduidige relatie tussen de uitkomstindicator en het zorgproces is niet beschreven. Of de relatie tussen het structuur of proceskenmerk en de zorguitkomsten is niet beschreven. Bovendien zijn er redenen om aan het bestaan van deze relatie te twijfelen.	De Hb1Ac- en bloeddrukscores bij type II diabetes patiënten zijn geen geschikte maat om de kwaliteit van een ziekenhuis te meten, omdat in vrijwel alle regio's in Nederland deze zorg met name in de eerste lijn wordt geleverd. Het te onderzoeken zorgproces (ziekenhuiszorg) heeft hier dan dus niet de gewenste, aantoonbare relatie met de uitkomsten van zorg. Het percentage trombose profylaxe bij patiënten in dagbehandeling is geen geschikte indicator, aangezien tromboseprofylaxe niet is geïndiceerd bij patiënten in dagbehandeling of bij patiënten die snel mobiel zijn.

3. Vertekening: populatievergelijkbaarheid

3.1. Definitie

Populatievergelijkbaarheid
De indicatorwaarden van verschillende zorgaanbieders op een bepaalde indicator zijn vergelijkbaar. Berekende indicatorwaarden weerspiegelen daadwerkelijke verschillen in de kwaliteit van de geleverde zorg en niet de verschillen in de populatiekarakteristieken van de zorgaanbieders.

Als zorguitkomsten bij gelijke kwaliteit zorg verschillen omdat de populatiekenmerken tussen zorgaanbieders verschillen, dan zijn de indicatorwaarden van zorgaanbieders niet vergelijkbaar. Om de zorgaanbieders toch te kunnen vergelijken moet correctie plaatsvinden voor populatiekenmerken. Indicatoren zijn vergelijkbaar indien

- correctie voor populatiekenmerken niet nodig is, *of*
- adequaat wordt gecorrigeerd voor verschillen in populatiekenmerken.

Populatiekenmerken verstoren de vergelijking tussen zorgaanbieders als de factoren voldoen aan minimaal de volgende drie criteria:

1. ze beïnvloeden de indicatorwaarde (kwaliteit van zorg);
2. ze zijn ongelijk verdeeld over zorgaanbieders; en
3. ze zijn niet door de zorgaanbieder te beïnvloeden.

Het populatiekenmerk leeftijd voldoet vaak aan deze criteria, omdat (1) leeftijd voor vele indicatoren een risicofactor is, (2) de gemiddelde leeftijd van de patiënten verschilt tussen de zorgaanbieders, en (3) de zorgaanbieder kan de leeftijd van een patiënt niet beïnvloeden. Populatiekenmerken verstoren vooral op de verschillen in 'zwaarte' van patiënten tussen zorgaanbieders. Het gaat daarbij om verschillende kenmerken van de behandelpopulatie, zoals ervaren gezondheid, leeftijd, geslacht, etniciteit, opleiding, sociaal-economische status, gesproken taal, etniciteit, ernst van somatische of psychiatrische aandoening, functionele status, co-morbiditeit, cognitief functioneren, psychosociaal functioneren, en attitudes en preferenties van de patiënten (Iezzoni 2003).

Per aandoening en zelfs per indicator kunnen de populatiekenmerken waarvoor gecorrigeerd moet worden verschillen. Voor patiëntervaringsindicatoren staat in de vragenlijst-specifieke werkinstructie per vragenlijst beschreven voor welke achtergrondkenmerken gecorrigeerd dient te worden. Patiëntkenmerken die bijvoorbeeld de kans op decubitus verhogen, hoeven niet per se dezelfde te zijn als patiëntkenmerken die de kans op sterfte na een hartinfarct verhogen. De populatiekenmerken zullen in de meeste gevallen medisch van aard zijn, maar sociaal-economische en culturele factoren komen ook in aanmerking. Indien patiënten uit lagere sociaal economische klassen in heel Nederland gemiddeld slechtere uitkomsten hebben, dan zullen zij waarschijnlijk bij gelijke kwaliteit van zorg toch slechtere uitkomsten hebben. In dat geval dient ook voor deze factoren te worden gecorrigeerd.

In theorie is het mogelijk om voor een grote hoeveelheid populatiekenmerken te corrigeren. Perfecte correctie bestaat echter niet. Het is zaak voor alle aspecten te corrigeren die voldoen aan de criteria en daarbij een significante invloed hebben op de onderlinge vergelijking van zorgaanbieders. De drie criteria voor een verstorend populatiekenmerk sluiten aan bij de

wetenschappelijke literatuur over epidemiologische methoden en prestatiemeting. De criteria zijn noodzakelijk maar niet voldoende om correctiefactoren te selecteren: alle correctiefactoren moeten aan de criteria voldoen, maar niet alle factoren die aan de criteria voldoen zijn factoren waarvoor dient te worden gecorrigeerd. Arbeidsmarktverschillen voldoen bijvoorbeeld aan alle criteria, maar een correctie voor arbeidsmarktverschillen zou verschillen in kwaliteit verhullen en is daardoor onwenselijk.

Populatiekenmerken voldoen echter wel. Zonder correctie zal bij gelijke kwaliteit de indicatorwaarde verschillen tussen zorgaanbieders met afwijkende populaties. Deze vertekening heet confounding. Dan kan de patiënt zijn keuze niet baseren op de indicatoren.

Voor een ander gebruikerdoel van de indicatorenwaarden dan keuze-informatie gelden andere eisen aan de selectie van correctiefactoren. Zo zullen zorgaanbieders voor interne verbeterinformatie bijvoorbeeld veelal geen behoefte hebben aan correctie (zie bijlage 3).

Indien een factor wel kan worden beïnvloed door de zorgaanbieder, dan is dat een mediërende factor waarvoor niet dient te worden gecorrigeerd. Indien een factor zowel deels bepaald is voor de behandeling en deels beïnvloed wordt tijdens de behandeling dan kan het best gevraagd worden naar de toestand aan het begin van de behandeling. Indien dat niet kan, dan dient een afweging gemaakt te worden tussen de voor- en nadelen van het corrigeren voor de factor. Bijlage 3 bevat een schematische toelichting op de begrippen.

3.2. Methoden om te corrigeren voor verschillen in populatiekenmerken

Er zijn verschillende technieken voorhanden om de populatievergelijkbaarheid van indicatoren te verhogen. Hieronder worden achtereenvolgens stratificatie en directe standaardisatie (Iezzoni 2003; Committee Performance Measures 2006) besproken. De methoden beantwoorden verschillende vragen en hebben verschillende statistische eigenschappen.

Stratificatie, indirecte standaardisatie en 'standaardisatie door middel van regressie' vergelijken alle de werkelijke uitkomst met de verwachtte uitkomst *voor de behandelde populatie*. Bij stratificatie gebeurd dit door homogene groep (strata) te maken op basis van de populatiekenmerken. Indien geslacht de enige propulatiefactor is dan zouden er twee strata worden aangemaakt: vrouwen en mannen. Vervolgens wordt de gemiddelde uitkomst van de zorgaanbieder per geslacht vergeleken met de gemiddelde uitkomst van datzelfde stratum van alle zorgaanbieders. Bij indirecte standaardisatie en regressie wordt per respondent de voorspelde uitkomst geschat op basis van de populatiekenmerken en wordt vervolgens het verwachtte gemiddelde uitgerekend. Deze technieken beantwoorden de vraag of een zorgaanbieder goede zorg heeft geleverd aan de patiënten die in zorg waren. Deze technieken zijn vooral bruikbaar om te beoordelen of een zorgaanbieder goede zorg levert voor de groep die wordt behandeld.

Directe standaardisatie en *inverse probability weighting* 'wegen' de patiënten in zorg zodanig dat haar karakteristieken overeenkomen met die van *een referentie populatie*, bijvoorbeeld de gemiddelde patiënten in zorg in Nederland. Indien mannen in Nederland 50% van de patiënten uitmaken, maar bij een zorgaanbieder slechts 25%, dan krijgen die mannen een gewicht van 2 en de vrouwen een gewicht van 2/3 waardoor de karakteristieken van de behandelpopulatie overeenkomt met het landelijk gemiddelde. Na deze weging zijn de uitkomsten direct vergelijkbaar tussen zorgaanbieders. Deze technieken zijn vooral bruikbaar voor het onderling vergelijken van zorgaanbieders voor selectieve zorginkoop.

In de praktijk is het onduidelijk of de verschillen tussen de beide technieken wezenlijk zijn. In theorie zijn de verwachte verschillen tussen beide technieken groot wanneer zorgaanbieders zich specialiseren en daardoor sterk afwijkende patiëntenpopulaties aantrekken. In de praktijk is dat mogelijk niet relevant omdat zorgaanbieders die zich op sterk verschillende populaties richten niet

direct zullen worden vergeleken bij een keuze door/voor patiënten (zie bijlage 6). Beide technieken zijn bruikbaar voor zowel relatieve als absolute uitkomstmaten (zie bijlage 7).

3.2.1. Stratificatie, indirecte standaardisatie en standaardisatie door regressie

In het geval van stratificatie wordt de patiëntenpopulatie opgesplitst in relatief homogene groepen. Correctie voor populatieverschillen wordt hier gerealiseerd door kwaliteit van zorg alleen binnen 'vergelijkbare groepen' patiënten te vergelijken (verschillende leeftijdsklassen, bijvoorbeeld, of alleen bij mannen). Stratificatie is het meest eenvoudig en wordt veelal direct begrepen door niet ingewijden, maar heeft als nadeel dat deze techniek veel waarnemingen vereist.

Bij zowel indirecte standaardisatie als standaardisatie door regressie wordt de verwachte uitkomst van een patiënt bepaald op basis van zijn kenmerken. Vervolgens wordt de verwachte uitkomst vergeleken met de werkelijke uitkomst. Voor adequate correctie zijn betrouwbare, gevalideerde prognostische modellen nodig die de kans op een uitkomst uitdrukken als functie van een reeks populatiekenmerken. Sommige van de genoemde populatiekenmerken, zoals leeftijd, co-morbiditeit, en socio-economische achtergrond, kunnen terugkomen bij verschillende indicatoren, maar zullen per indicator verschillen in hun effect.

Indirecte standaardisatie is als concept eenvoudiger dan standaardisatie door regressie. Regressie is echter gemakkelijker uit te voeren. In de praktijk is het belangrijkste voordeel van regressie dat de techniek efficiënt is. Dat wil zeggen dat de techniek een lager aantal waarnemingen vereist om verschillen tussen aanbieders als statistisch significant te kunnen duiden.

De regressietechniek vereist hiervoor wel een extra veronderstelling: de populatiekenmerken mogen *niet allen* interacteren[9]. Dat wil zeggen, het effect van een populatiefactor mag niet systematisch samenhangen met de waarde van andere populatiekenmerken. Gesteld dat geslacht en leeftijd populatiekenmerken zijn, dan mag het verschil tussen mannen en vrouwen op de uitkomst niet samenhangen met leeftijd. Deze veronderstelling dient te worden getest. Indien de veronderstelling verworpen wordt kan het regressiemodel worden aangepast, maar dan zal het regressiemodel vergelijkbaar zijn met indirecte standaardisatie en evenveel waarnemingen nodig hebben als indirecte standaardisatie.

3.2.2. Directe standaardisatie en inverse probability weighting

Bij zowel directe standaardisatie als inverse probability weighting (IPW) wordt de patiëntenpopulatie gewogen zodat de populatie na weging vergelijkbaar is met een referentiepopulatie, bijvoorbeeld de Nederlandse bevolking. Indien geslacht de enige populatiefactor is dan krijgen alle vrouwen en alle mannen in de patiëntenpopulatie per geslacht een verschillend gewicht. De IPW-gewichten kunnen met regressie worden geschat en daardoor heeft IPW minder waarnemingen nodig dan directe standaardisatie. IPW vereist ten opzicht van directe standaardisatie aanvullende modelaannames die kunnen worden getoetst.

Voor een succesvolle standaardisatie moeten de populatiekenmerken ook op patiëntniveau geregistreerd zijn. Voor standaardisatie is informatie nodig over allerlei factoren die de variantie in de indicator mede verklaren (leeftijd, geslacht, etc.). Hoe specifieker de data zijn, hoe hoger de kwaliteit van het regressiemodel en daarmee van de populatiecorrectie. Indien de data niet op individueel maar op geaggregeerd niveau worden verzameld leidt dat tot drie oorzaken van vertekening:

1. correctie voor populatieverschillen zal onvoldoende zijn als gevolg van vervlakkingsvertekening (attenuation bias);
2. correctie voor interacties tussen populatiekenmerken is niet mogelijk;

[9] Indien alle factoren wel interacteren en opgenomen worden in het regressiemodel, dan spreken we van een verzadigd (saturated) regressiemodel.

3. correctie levert het risico op van aggregatievertekening (ecological fallacy), omdat 'gemiddelde' gegevens van groepen weinig zeggen over de relatie tussen zorzwaartefactoren en de uitkomsten tussen individuën.

Het is hierdoor op voorhand onduidelijk of een correctie op basis van deze geaggregeerde data de vergelijkbaarheid laat toe- of afnemen.

3.3. Populatievergelijkbaarheid in relatie tot het type indicator

De populatievergelijkbaarheid van een *structuurindicator* is meestal goed. Het hebben van decubitusbeleid of een veiligheidssysteem is immers onafhankelijk van de kenmerken van de behandelpopulatie van een zorgaanbieder. In dit geval is het toetsen van de populatievergelijkbaarheid niet nodig.

Voor sommige *procesindicatoren* gaat deze argumentatie eveneens op. In het geval van de indicator 'het percentage rokende patiënten met diabetes waarbij stoppen met roken advies is gegeven', kan ook worden gesteld dat dit niet aan de populatie is gekoppeld. Er wordt immers alleen gemeten of het advies is gegeven – niet of het ook is opgevolgd. Patiëntkenmerken behoren dan geen rol te spelen en een beoordeling van de vergelijkbaarheid is niet nodig. In het geval van doorlooptijden of medicatie-compliantie vragen (heeft de patiënt de adequate medicatie gekregen na zijn myocard infarct, bijvoorbeeld) kan de behandelpopulatie wel relevant zijn. Patiënten met ernstige co-morbiditeit zullen bijvoorbeeld niet altijd de standaard medicatie kunnen verdragen.

Voor *uitkomstindicatoren* ten slotte geldt dat de populatievergelijkbaarheid vrijwel altijd relevant is. Bij een indicator als 'percentage patiënten met decubitus' of '5-jaars overleving na behandeling voor borstkanker' is de correctie voor populatieverschillen van groot belang: uitkomsten zijn immers niet alleen van het zorgproces zelf afhankelijk, maar ook van de betreffende patiënten. Patiëntervaringen worden vaak gecorrigeerd voor ervaren gezondheid, opleiding en leeftijd. Ouderen zijn over het algemeen positiever over de zorg dan jongeren, mensen in goede gezondheid positiever dan mensen in slechte gezondheid, en laag opgeleiden positiever dan hoog opgeleiden.

3.4. Beoordelen van populatievergelijkbaarheid

Aan indicatoren waarvoor de populatievergelijkbaarheid relevant is dient een vlag te worden toegekend met behulp van de onderstaande tabel. Voor structuurindicatoren en een deel van procesindicatoren is dit dus niet nodig. In deze tabel staat beschreven welke signaalvlaggen voor het criterium populatievergelijkbaarheid kunnen worden toegekend.

Vlag	Toelichting	Voorbeeld
Voldoet	Een populatiecorrectie methodiek wordt gehanteerd waarbij de populatiekenmerken voor correctie goed zijn onderbouwd. *Of* Er is onderbouwd dat er geen verschillen in populatie tussen zorgaanbieders zijn die de indicatorwaarde significant beïnvloeden.	Bij de indicator 'Huidletsel' (VV&T) wordt gecorrigeerd voor verschillende populatiekenmerken, waaronder elementen uit de Care Dependency Scale (zie leeswijzer bij het Prestatie-overzicht Verantwoorde Zorg VV&T, v 1.6). Deze correctie vindt plaats o.b.v. data op patiëntniveau. Bij bijvoorbeeld de volume-indicatoren in de ziekenhuiszorg is populatie-correctie niet relevant: eventuele verschillen in populatie zijn op zichzelf niet belangrijk voor de vraag hoeveel ervaring een zorgaanbieder heeft met een bepaalde interventie. (Als er veel ervaring is, kan de vraag naar

Vlag	Toelichting	Voorbeeld
		eventuele subspecialisatie wel weer relevant worden).
Voldoet deels	Indien bij een procesindicator zonder adequate onderbouwing wordt afgezien van correctie voor populatiekenmerken. *Of* de populatiecorrectie die wordt gehanteerd neemt niet de relevante populatiekenmerken mee of is anderszins beperkt onderbouwd.	Zie indicator 3a set cataract van ziekenhuizen. De tijdsperiode tussen operatie van 1e en 2e oog ≥ 28 dagen. De noodzaak van populatiecorrectie lijkt voor deze procesindicator wel aanwezig. In de factsheet staan verschillende vormen van comorbiditeit beschreven, waarvoor afweken kan worden van de grens van > 28 dagen. Voor een goede vergelijking van ziekenhuizen moet worden onderzocht van de invloed is van comorbiditeit en indien relevant moet een model voor populatiecorrectie voorhanden zijn.
Voldoet niet	Indien bij een uitkomstindicator zonder onderbouwing wordt afgezien van correctie voor populatiekenmerken. *Of* indien bij een uitkomst- of procesindicator het op basis van literatuur of expert opinion relevant wordt geacht een correctie uit te voeren, maar hierin niet of marginaal is voorzien.	Voor de basisset prestatie indicatoren GGZ wordt geen populatiecorrectie uitgevoerd bij uitkomst- en procesindicatoren (m.u.v. een aantal indicatoren waarbij stratificatie naar diagnosegroepen plaats vindt). Experts hebben echter aangegeven dat populatiecorrectie wenselijk is om de populatievergelijkbaarheid van de uitkomsten van de indicatoren te bevorderen (ZiZo, 2009f). Bijvoorbeeld een correctie voor de ernst van een aandoening of het wel of niet hebben van een comorbide stoornis bij indicator 'Verandering ernst problematiek'. De invloed van deze factoren is echter nog niet statistisch onderbouwd.

4. Vertekening: registratievergelijkbaarheid

4.1. Definitie

Registratievergelijkbaarheid

Registratievergelijkbaarheid ontstaat wanneer vergelijkbaarheid van indicatoren niet wordt aangetast door verschillen tussen de wekelijke en de geregistreerde waarden van structuur-, proces-, uitkomst- en populatiewaarden.

Met registratievergelijkbaarheid wordt gerefereerd aan het proces van meten, registreren en aanleveren van de benodigde gegevens. Dit dient juist te gebeuren zonder meetfouten omdat de gegevens anders niet overeenstemmen met wat werkelijk is gebeurd. Deze vertekening werd in een eerdere versie van de indicatorstandaard registratiebetrouwbaarheid genoemd.

Vrijwel geen enkele meting of registratie is vrij van meetfouten. Meetfouten kunnen betrekking hebben op de variabelen van de indicatoren zelf, maar ook van de variabelen die voor een eventuele populatiecorrectie worden gebruikt. Meetfouten kunnen het gevolg zijn van toeval of slordigheid, maar er kan ook sprake zijn van strategische invulling van indicatorwaarden of populatievariabelen. In het laatste geval leidt dat vanzelfsprekend tot vertekening van de indicatoruitkomsten, waarbij een zorgaanbieder die de indicatorwaarden of de populatievariabelen strategisch invult in het voordeel is.

Als het gaat om de populatievariabelen kunnen ook meetfouten die het gevolg zijn van toeval of slordigheid tot vertekening leiden. Dit is de zogenaamde vervlakkingsvertekening (attenuation bias)[10]: dergelijke meetfouten in de meting van populatievariabelen leiden tot een onderschatting van daadwerkelijke effecten van zorgzwaartefactoren op de indicatorwaarde. Dit leidt tot ondercorrectie voor populatieverschillen, waardoor zorgaanbieders met een lichte behandelpopulatie bij gelijke kwaliteit betere uitkomsten krijgen. Bij een onderlinge vergelijking zullen de zorgaanbieders met relatief gezonde patiënten onterecht beter scoren dan zorgaanbieders met ongezonde complexe patiënten. In Bijlage 4 worden de verschillende typen meetfouten toegelicht met schematische figuren en voorbeelden.

Een indicator met een hoge registratievergelijkbaarheid heeft relatief weinig last van meetfouten. Eventuele meetfouten kunnen worden voorkomen door een aantal maatregelen:

- de definitie van de indicator dient scherp te zijn;
- de indicatorspecificaties dienen te borgen dat registraties en aanlevering tijdig, en juist geschieden;
- de indicator baseert zich op geschikte databronnen;
- de indicatorgegevens dienen retrospectief controleerbaar te zijn (ZiZo 2009d).

4.2. Methoden om registratievergelijkbaarheid te verbeteren

Er zijn meerdere methoden om te komen tot een goede registratie van indicatoren. De meest gebuikelijke worden hieronder besproken.

[10] In de wetenschappelijke literatuur staat deze vorm van bias bekend als *attentuation bias, regression dilution, (classical) errors-in-variables*

4.2.1. Scherpe definities

Een goede indicator kan niet zonder goede definities. Alle essentiële componenten van de indicator dienen helder, precies en volledig te zijn omschreven. Dit betekent dat scherp is vastgelegd:

- wat de aard en omvang van de aandoeningen en/of zorgproces is waar de indicator betrekking op heeft;
- wat het organisatorisch verband is waar de indicator betrekking op heeft (bijv. zorg door een verpleeghuis, de thuiszorg, een individuele specialist, een afdeling, of ketenzorg);
- wat wordt gemeten (scherpe definitie van het fenomeen, omschrijving van de teller);
- wat de doelpopulatie is (in- en exclusiecriteria zijn scherp; de noemer is scherp is omschreven);
- wat de definities zijn van gehanteerde terminologie in de teller en de noemer (de teller dient altijd een deelverzameling van de noemer te zijn);
- wat het meetinstrument is [11];
- hoe de data dienen te worden geregistreerd, verwerkt en aangeleverd;
- wat voor een type uitkomsten worden gepresenteerd (dichotome, ordinale of intervalschaal-uitkomsten);
- wat de termijnen zijn waar de indicator betrekking op heeft (wanneer begint en eindigt de meetperiode, wat valt er nog net binnen, wat niet).

In alle gevallen dienen relevante bijzondere situaties en uitzonderingen te zijn beschreven.

Indien wordt gewerkt met een meetinstrument (een vragenlijst of bijvoorbeeld een pijnscore-schaal of meetapparaat) dient dat instrument *zelf* ook gevalideerd te zijn (dit type validering wordt hier verder niet besproken; zie bijvoorbeeld de handleiding van het CKZ over de validering van vragenlijsten)[12]. Hoe met de vragenlijst omgegaan dient te worden staat voor de patiëntervaringsindicatoren met de CQ-index methodiekbeschreven in de specifieke werkinstructie van die vragenlijst. Deze is aanvullend op het Handboek Metingen waarin algemene regels zijn opgenomen zoals hoe een indicator berekend dient te worden[13].

Indien meerdere meetinstrumenten worden toegestaan voor dezelfde indicator dan dient aangetoond te zijn dat de verschillende meetinstrumenten in dezelfde omstandigheden vergelijkbare meetresultaten opleveren. Die vergelijkbaarheid is mede afhankelijk van de wijze waarop de resultaten vergelijkbaar worden gemaakt.
Wanneer verschillende meetinstrumenten voor een indicator verschillende vragen hanteren, dan leidt dit veelal ook tot afwijkende uitkomstmaten. Daardoor zijn de uitkomsten tussen de beide indicatoren niet direct vergelijkbaar. De uitkomsten kunnen vergelijkbaar worden gemaakt, mits de indicatoren hetzelfde beogen te meten. Twee technieken zijn hiervoor voor de hand liggend:

1. Beide indicatoren transformeren tot een z-schaal (http://en.wikipedia.org/wiki/Standard_score). Bij deze techniek wordt van elke waarde het gemiddelde afgetrokken en wordt het residu gedeeld door de standaarddeviatie. Hierdoor krijgt elke indicator een z-score toebedeeld en die zijn dan tussen de registratiesystemen vergelijkbaar.
2. Een van de twee indicatoren uitdrukken in de schaal van de ander. Voor deze techniek wordt het verschil in gemiddelde tussen de twee indicatoren berekend. Het verschil wordt opgeteld bij de waarden van de verdeling die wordt uitgedrukt in de andere. Hierdoor zijn de gemiddelden gelijk. Vervolgens wordt de standaarddeviatie op vergelijkbare wijze aangepast.

[11] Of de teller en noemer *juist* zijn gedefinieerd, en het meetinstrument past bij de doelstelling van de indicator is een vraag naar de inhoudsvaliditeit van de indicator: zie hoofdstuk 3. Bij de selectie van de juiste noemer (de doelpopulatie) spelen ook vergelijkbaarheidsoverwegingen een rol: zie hoofdstuk 4.
[12] Andere voorbeelden zijn de Care Dependency Scale, de SNAQ (voeding), ASA score .
[13] zie WIS 07.01 Werkinstructie berekening CQI-schaalscores, Handboek CQI Metingen

4.2.2. Juiste vastlegging

De indicatorinstructies dienen precies aan te geven hoe (en zo nodig door wie) de data dienen te worden geregistreerd en aangeleverd. Op dit moment worden zorginhoudelijke gegevens nog grotendeels achteraf en 'met de hand' uit (vaak incomplete) papieren zorgdossiers gehaald (Kallewaard et al. 2007). Dit proces is zeer gevoelig voor fouten, al is het alleen maar omdat veel interpretatie nodig is om de zorggegevens te vertalen in de voor de indicatoren benodigde gegevens. Een dergelijke werkwijze levert bovendien een hoge administratie lastendruk (Zichtbare Zorg 2008).

In de toekomst zullen zorgaanbieders de benodigde data steeds meer prospectief en automatisch gaan verzamelen, als onderdeel van het eigen zorgproces en de sturing daarvan. Dit zal de uniformering en daarmee juistheid van de registraties ten goede komen. Dit geldt zowel voor de indicatorwaarden als populatievariabelen.

Bij klinische observaties of andere registraties waarbij een vorm van interpretatie onontkoombaar is, is het essentieel om ook uitspraken te doen over de benodigde registratie-expertise. Moet een arts of medisch student de registraties doen, mogen andere professionals de data ook invoeren, of mag alleen een speciaal getrainde codeur registreren? Voor de landelijke decubitus metingen van de Universiteit Maastricht wordt veel aandacht besteed aan het trainen en faciliteren van de personen die de metingen verrichten. Om de juistheid van de populatievariabelen te bevorderen zijn ook uniforme landelijke definities nodig.

De juiste vastlegging van de data die via vragenlijsten worden verzameld dient te zijn geborgd via meetinstructies van de betreffende vragenlijsten.

4.2.3. Geschikte databronnen

De indicatorinstructies dienen ook precies aan te geven welke databronnen dienen te worden gehanteerd. Die keuze heeft grote invloed op de kans op meetfouten.

Bij indicatoren over patiëntervaringen zijn patiënten zelf de informatiebron. Door middel van schriftelijke en online vragenlijsten of interviews worden gegevens verzameld. De kans op meetfouten wordt vooral bepaald door de kwaliteit en de uitvoering van de meetinstructies.

Voor zorginhoudelijke indicatoren zijn bij zorgaanbieders verschillende bronnen beschikbaar waarin patiëntgegevens zijn vastgelegd: naast administratieve systemen ook klinische gegevens in patiëntendossiers. De geschiktheid van deze bronnen varieert. De meest wenselijke werkwijze is dat administratieve gegevens en basale proces parameters uit administratieve systemen worden afgeleid, en klinische gegevens in het primaire proces van zorg zelf worden vastgelegd. Als dit secuur en elektronisch gebeurt, op basis van landelijke standaarden, dan zijn deze data in principe de perfecte bron voor het automatisch genereren van de gevraagde indicatorwaarden. Als een dergelijke registratie structureel in het reguliere zorgproces is ingebed spreken we van een zorginhoudelijke registratie. Goed ingebed kost zo'n registratie in vergelijking met post-hoc dataverzameling minder werk.

De hierboven genoemde 'perfecte' situatie voor wat betreft de *klinische* gegevens wordt benaderd in regionale of landelijke registraties die expliciet zijn opgezet om informatie voor zorgvergelijkingen te genereren. Kenmerk van al deze zorginhoudelijke registraties is dat specifiek aandacht wordt besteed aan de kwaliteit van de dataregistratie en aan de juistheid van de data. De secuurheid van de registraties is geborgd via procedures of kwaliteitscontroles, waaronder autorisatie door de verantwoordelijke professional indien die professional niet zelf de gegevens heeft vastgelegd. Vanuit regionaal of landelijk niveau worden de indicatoren van individuele zorgaanbieders actief teruggekoppeld via benchmarks. Door de aandacht voor en het monitoren

van de kwaliteit van de data en het actief gebruik van die data in kwaliteitsmanagement is de betrouwbaarheid van dit type data hoog.

Indicatoren kunnen ook worden gebaseerd op *administratieve* gegevens. Dit zijn gegevens die zorgaanbieders standaard vastleggen vanwege bijvoorbeeld declaratie-eisen: DBC gegevens, opnamegegevens, ANW gegevens, medicatiegegevens, gegevens over de verrichte diagnostische en therapeutische activiteiten, enzovoort. Ondanks een kleinere kans op verwerkingfouten en de mogelijkheid om de opgeleverde gegevens te controleren varieert de juistheid van dit type gegevens toch wezenlijk (Van den Bosch, 2011)

4.2.4. Retrospectieve controleerbaarheid

Bij het gebruik van interne databronnen dient de vastlegging dusdanig te zijn dat het voor een interne of een externe auditor mogelijk is om een oordeel te geven over de correcte vastlegging van de gegevens, bijvoorbeeld door middel van een steekproefcontrole. Hiertoe dient vastgelegd te zijn welke patiënten meetellen in de doelpopulatie van een specifiek jaar, welke patiënten geëxcludeerd zijn (en waarom). Daarnaast dienen de gegevens per patiënt in de primaire systemen traceerbaar te zijn, inclusief gelogde informatie over datum van vastlegging, wijzigingen, auteur, enzovoort.

4.3. Registratievergelijkbaarheid in relatie tot het type indicator

Voor *structuurindicatoren* ('is er een systematische decubitusregistratie') geldt dat de meeste eisen voor registratievergelijkbaarheid niet relevant zijn: er hoeft immers alleen te worden aangegeven of er wel of niet een bepaald structuurkenmerk aanwezig is. Ook de controleerbaarheid is in het geval van de aan- of afwezigheid van een structuurkenmerk zelden complex. Dat betekent dat de registratievergelijkbaarheid van een structuurindicator vooral afhangt van de volledigheid en compleetheid van de gehanteerde definities. Als de criteria om 'ja' of 'nee' te scoren op een vraag over de aanwezigheid van een structuurkenmerk niet precies zijn gedefinieerd, dan blijft de registratievergelijkbaarheid onder de maat.

Voor *procesindicatoren* zijn de criteria voor registratievergelijkbaarheid relevanter. Bij een procesindicator 'het percentage patiënten dat wordt gescreend op ondervoeding' is bijvoorbeeld niet alleen de definitie van belang, maar ook de wijze van registratie en berekening.

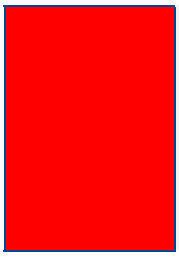
Voor *uitkomstindicatoren* tenslotte is de registratievergelijkbaarheid uiterst relevant. Bij een indicator 'percentage patiënten met op enig moment een pijnscore boven de 7 in de eerste 72 uur na een operatie' zijn zowel definitie als registratie en berekenmethode van groot belang.

4.4. Beoordelen registratievergelijkbaarheid

In de onderstaande tabel staat beschreven, en met voorbeelden geïllustreerd, hoe de signaalvlaggen groen, geel en rood voor het criterium registratievergelijkbaarheid toegekend kunnen worden. De registratievergelijkbaarheid van een indicator wordt in zijn geheel beoordeeld. Er wordt dus geen onderscheid gemaakt naar tellers en noemers van indicatoren.

Vlag	Toelichting	Voorbeeld
Voldoet	De indicator is helder, precies en volledig omschreven, inclusief beschrijving van *wat* er wordt gemeten, de in- en exclusiecriteria van de doelpopulatie, de teller/noemer (indien relevant) en het meet-	De indicator "Het totaal aantal cystectomieën" uit de indicatorset Blaascarcinoom is voorzien van coderingen voor zorgactiviteiten. De vraagstelling en bijbehorende definities zijn eenduidig. En de gegevens worden in de administratieve gegevens

Vlag	Toelichting	Voorbeeld
	instrument (indien relevant). De meet-, registratie- berekenings- en rapportage instructies zijn volledig en precies omschreven. In het geval dat een meetinstrument wordt gebruikt gaat het om een gevalideerd instrument. Indien meerdere meetinstrumenten zijn toegestaan is aangetoond dat de meetinstrumenten dezelfde uitkomsten genereren in dezelfde omstandigheden. Data worden (elektronisch en) eenduidig op patiëntniveau vastgelegd, idealiter tijdens het zorgproces zelf. De tijdigheid, correctheid en volledigheid van deze vastlegging is geborgd via procedures of kwaliteits-controles, waaronder autorisatie door de verantwoordelijke professional indien die niet zelf de gegevens heeft vastgelegd. Ook zijn de vastgelegde data retrospectief controleerbaar (op patiëntniveau), inclusief gelogde informatie over datum van vastlegging, wijzigingen, auteur, enzovoort. Er is, met andere woorden, minimale ruimte voor datamanipulatie.	vastgelegd en kunnen geautomatiseerd worden afgeleid. Gegevens verzameld met een patiëntervaringsvragenlijst door een meetbureau waarbij in de specifieke werkinstructies de wijze van afname en in en exclusiecriteria beschreven zijn.
Voldoet deels	De indicator en bijbehorende definities en instructies zijn niet scherp omschreven en de indicator is voor meerdere interpretaties vatbaar. *Of* Data worden elektronisch vastgelegd, en worden door zorgaanbieders zelf gebruikt voor interne verbetering en sturing. Er is echter geen sprake van borging van de tijdigheid, correctheid en volle-digheid van de registratie door procedures of kwaliteitscontro-les. Eventuele datamanipulatie is niet uit te sluiten.	Het percentage reumatoïde artritis patiënten waarbij de ziekteactiviteit (DAS-28 score) binnen de afgesloten DBC tenminste éénmaal gemeten is. De indicator is zodanig geoperationaliseerd (incl. definities) dat de indicator niet voor meerdere interpretaties vatbaar is. De gegevens worden in de dagelijkse praktijk geregistreerd, maar kunnen niet geautomatiseerd door de ziekenhuizen uit de administratieve gegevens worden afgeleid. Noch is de kwaliteit van de klinische registratie van de indicatorwaarde geborgd.
Voldoet niet	De indicator is niet scherp omschreven, doelpopulaties zijn niet scherp afgebakend of het meetinstrument is niet	In de basisset prestatie indicatoren GGZ hebben zorgaanbieders de ruimte om verschillende meetinstrumenten te

Vlag	Toelichting	Voorbeeld
	gespecificeerd of niet gevalideerd. *Of* Data zijn niet correct of kunnen niet retrospectief worden gecontroleerd op patiëntniveau.	gebruiken, waarbij niet is aangetoond dat de meetinstrumenten dezelfde uitkomsten genereren in dezelfde omstandigheden. Dit geldt voor de indicatoren verandering ernst problematiek en verandering in de ervaren kwaliteit van leven.

5. Vertekening: steekproef- en responsvergelijkbaarheid

5.1. Definitie

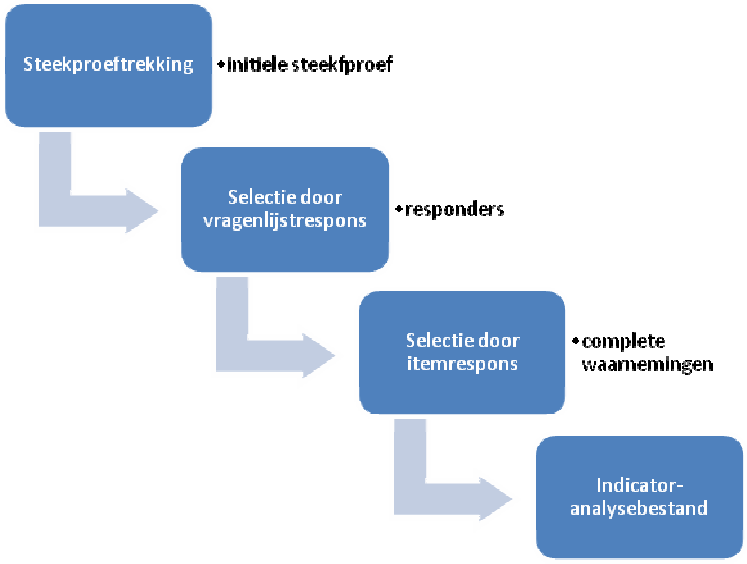
Steekproef- en responsvergelijkbaarheid

De vergelijkbaarheid van indicatorwaarden wordt niet beïnvloed door verschillen tussen de steekproef en repons ten opzichte van de totale behandelde populatie. Met andere woorden: de steekproef en respons is gezamenlijk representatief voor de behandelpopulatie.

In sommige gevallen is volledige registratie niet wenselijk vanwege de kosten, bijvoorbeeld bij zorgaanbieders met een groot aantal patiënten of met patiënten die langdurig worden verzorgd. In zo'n situatie kan een goede steekproef uitkomst bieden. Een goede steekproef is zowel representatief als voldoende groot. Dit hoofdstuk gaat over de representativiteit van de steekproef en de daaropvolgende respons. De steekproefgrootte komt in het volgende hoofdstuk aan bod.

Een steekproef bestaat uit een deel van alle behandelde patiënten. Vervolgens is de respons dat deel van de patienten waarover gegevens zijn verzameld voor de bepaling van een indicatorwaarde. Respons wordt onderverdeeld in vragenlijstrespons en itemrespons. Alleen de complete waarnemeningen worden opgenomen in het indicatoranalysebestand en gebruikt om de indicatorwaarden mee te bepalen, zie figuur 5.1.

Figuur 5.1 Van steekproeftrekking en respons naar indicatoranalysebestand



Men spreekt in het algemeen van verminderde representativiteit wanneer het selectief includeren van personen in een onderzoek leidt tot onvergelijkbare indicatorwaarden. Deze vertekening treedt op wanneer de selectie van waarnemingen zowel samenhangt met de zorgaanbieder, als de uitkomst van zorg.

Het effect van vertekening op indicatorvergelijkingen is dat verschillen tussen zorgaanbieders gevonden kunnen worden die er in werkelijkheid niet zijn. Het is ook mogelijk dat

steekproefvertekening echte verschillen in geleverde zorg verdoezelt of versterkt. Doordat de vertekening geen richting heeft, is de interpretatie van de vertekende resultaten erg lastig. Het belang van representatieve datasets is voor is daarom groot.

Voorbeelden steekproefvertekening

Verzorgingshuis B heeft moeite de personele bezetting tijdens vakantieperiodes op een goed niveau te houden. Het verpleeghuis kiest een meetweek in een periode waarin de bezetting wel goed is. Hierdoor is de geschatte kwaliteit hoger dan de werkelijk geboden kwaliteit in dat jaar.

Zorgaanbieder A bejegent haar patiënten slecht. Slecht behandelde patiënten vullen de patiëntervaringsvragenlijst veelal niet in. Hierdoor is de geschatte kwaliteit van zorg van A hoger dan de werkelijke kwaliteit van zorg.

Verpleeghuis A en B leveren beide gelijke kwaliteit van zorg. Verpleeghuis A vraagt ernstig zieke patiënten niet om deel te nemen aan de meting, omdat het huis hen niet wil belasten met het onderzoek. Verpleeghuis B probeert juist zoveel mogelijk patiënten bij de meting te betrekken om een zo compleet mogelijk beeld te krijgen. Bij gebruik van de indicator Decubitus zal verpleeghuis A in de indicatorvergelijking beter scoren dan verpleeghuis B.

Steekproefvergelijkbaarheid is gegarandeerd voor indicatoren die zijn gebaseerd op een registratie waarin alle behandelde patiënten volledig zijn opgenomen. Er is dan geen sprake van een steekproef of non-respons.

Uitspraken die gebaseerd zijn op een steekproef kunnen gevoelig zijn voor toevalsvariatie (zie hoofdstuk 6) en voor gebrek aan steekproefvergelijkbaarheid. Ten behoeve van het statistisch betrouwbaar onderscheiden (zie hoofdstuk 6) is het in veel gevallen wenselijk om voor de indicatoren relevante gegevens van alle patiënten gestandaardiseerd en continu vast te leggen. Dit verhindert ook vertekening door een niet-representatieve selectie van waarnemingen.

5.2. Methoden om steekproef- en responsvergelijkbaarheid te verbeteren

De steekproefvergelijkbaarheid is afhankelijk van de methodiek waarmee de initiële steekproef is getrokken: de steekproeftrekking. Goede vergelijkbaarheid kan worden gestimuleerd door de steekproeftrekking te laten voldoen aan twee voorwaarden:

1. De steekproeftrekking garandeert een representatieve steekproef. Deze voorwaarde pleit voor een meting zonder gekozen meetperioden en het optimaliseren van de respons.
    - *Meetperiode*: Een meetperiode kan leiden tot een steekproefvertekening doordat de zorg van sommige aanbieders sterker wisselt over de tijd dan bij andere aanbieders, bijvoorbeeld als gevolg van een slechte vakantieplanning. Hierdoor geven de indicatorwaarden geen getrouw beeld van de kwaliteit van zorgaanbieders ten opzichte van elkaar over de gehele periode waarover verslag wordt gedaan (bijvoorbeeld een verslagjaar).
    - *Selectieve respons*: Ook bestaat de kans dat specifieke groepen patiënten meer geneigd zijn dan andere groepen patiënten om vragenlijsten in te vullen en op te sturen. Bij uitkomstindicatoren en een deel van de procesindicatoren zijn populatiekenmerken (case-mix) gegevens nodig voor de correctie. Ook daar speelt uitval een rol. In een aantal gevallen is een selectieve steekproef niet te

voorkomen. Dit is bijvoorbeeld het geval wanneer de patiënten zelf beslissen of zij een vragenlijst zullen invullen. In dergelijke situaties kan de responsvergelijkbaarheid van de schatting worden verhoogd door te streven naar het maximaliseren van het aantal volledig ingevulde vragenlijsten en door kenmerken te verzamelen van alle patiënten die een vragenlijst hebben ontvangen. Deze kenmerken kunnen vervolgens worden gebruikt in de statistische correctie. Correctie voor een selectieve steekproef is echter slechts mogelijk wanneer de oorzaak van de selectie bekend en meetbaar is.

2. Zorgaanbieders hebben geen invloed op de selectie en respons van patiënt.
    - *Steekproeftrekking*: Een steekproeftrekking uit een bestand met alle behandelde patiënten kan representatief zijn, maar kan ook gevoelig zijn voor beïnvloeding. Zo kan een random procedure meerdere malen worden herhaald totdat een strategisch gunstige selectie van patiënten wordt getrokken. Daarnaast is een trekking op basis van toeval veelal inefficient, waardoor meer patiënten dienen te worden uitgevraagd dan bij een meer geavanceerde procedure.
    - *Definitie inclusiecriteria*: Inclusiecriteria kunnen de steekproef- en responsvergelijkbaarheid beïnvloeden indien deze verschillend worden geïnterpreteerd door zorgaanbieders, meetbureau's, of patiënten.
    - *Complete waarnemingen:* Indien zorgaanbieders de waarnemingen zelf registreren dan leveren zij voor de door de steekproef geselecteerde patienten volledige gegevens aan.

Om tot een goede representativiteit te komen is het noodzakelijk om een goede steekproefprocedure te ontwikkelen per sector, die zowel ongevoelig is voor manipulatie als efficient. Het ontwikkelen van zo'n procedure kan het best aan onderzoeksbureau's worden overgelaten. Elke partij kan de procedure uitvoeren, mits een controle mogelijk is waarmee de uiteindelijke steekproef kan worden gerepliceerd (dezelfde selectie van patiënten). Alternatieven zijn dat de steekproeftrekkingen extern worden uitgevoerd of in detail worden beschreven zodat controle van de procedure achteraf mogelijk is. Toetsing van vertekening achteraf is lastig, omdat geregistreerde gegevens van patiënten veelal niet zullen volstaan om selectieve steekproeftrekking vast te stellen.

Kostenoverwegingen zijn veelal de aanleiding om met steekproeven te werken. De efficiëntie, en daarmee de kosten, van een steekproef kunnen verder worden verbeterd door gebruik te maken van steekproeftrekking-technieken, zoals *oversampling* op basis van verwachte variatie, *stratified sampling*[14] of *multi-stage sampling*.

5.3. Steekproef- en responsvergelijkbaarheid in relatie tot het type indicator

Respresentativiteit van de data speelt bij structuurindicatoren en procesindicatoren die op zorgaanbiederniveau worden uitgevraagd geen rol omdat de resultaten niet afhankelijk zijn van een steekproef of respons. Indicatoren die op patiëntniveau worden uitgevraagd zijn wel gevoelig voor de selectie van patiënten waar de indicator op is gebaseerd. Voorbeelden hiervan zijn alle zorginhoudelijke uitkomstindicatoren (waaronder PROM's), procesindicatoren waarvan de score verschilt tussen patiënten en uitkomstindicatoren op basis van patiëntervaringsvragenlijsten.

[14] Initiele steekproeven worden bij de CQ-index methodiek vaak gecontroleerd op hun representativiteit ten aanzien van enkele factoren (bijvoorbeeld leeftijd en geslacht). Indien de verdeling in de steekproef afwijkt van de populatie, dan wordt er een nieuwe steekproef getrokken. Deze benadering kan worden beschouwd als een vorm van stratified sampling.

5.4. Beoordelen steekproefvergelijkbaarheid

In de onderstaande tabel staat beschreven, en met voorbeelden geïllustreerd, hoe de signaalvlaggen groen, geel en rood voor het criterium steekproefvergelijkbaarheid toegekend kunnen worden. De steekproefvergelijkbaarheid van een set indicatoren wordt in zijn geheel beoordeeld. Er wordt dus geen onderscheid gemaakt naar indicatoren.

Vlag	Toelichting	Voorbeeld
Voldoet	De indicator is gebaseerd op een volledige registratie van alle patiënten. *Of* De procedure voor de steekproeftrekking is beschreven en achteraf eenduidig reproduceerbaar, en van (vrijwel) alle patiënten uit de steekproef worden volledige gegevens aangeleverd door de zorgaanbieder. *Of* De procedure voor de steekproeftrekking is beschreven en achteraf eenduidig reproduceerbaar, en ten minste van 55% van alle patiënten uit de steekproef zijn volledige gegevens beschikbaar. *Of* De procedure voor de steekproeftrekking is beschreven en achteraf eenduidig reproduceerbaar, en ten minste 30% van alle patiënten uit de steekproef leveren volledige gegevens aan, en er is een statistische correctie uitgevoerd.	Een ziekenhuisindicator voor diepe wond infecties waarbij volledige gegevens van alle behandelde patiënten van het meetjaar door alle ziekenhuizen worden aangeleverd Een indicator op basis van een patiëntervaringsmeting waarbij het respons percentage gelijk of hoger is dan 55% en de respondenten aangetoond representatief zijn voor de populatie van de zorgaanbieder.
Voldoet deels	De procedure voor de steekproeftrekking is beschreven, maar is achteraf *niet* eenduidig reproduceerbaar (het doorlopen van de procedure kan leiden tot een afwijkende steekproef). *Of* Er wordt gebruik gemaakt van een meetperiode die korter is dan de periode waarover verslag wordt gedaan, en over alle patiënten worden gegevens	Een patiëntervaringsmeting waarbij het respons percentage is minder dan 30% en de respondenten zijn representatief zijn voor de populatie van de zorgaanbieder of er is een statitische correctie toegepast i.

Vlag	Toelichting	Voorbeeld
	aangeleverd. *Of* De procedure voor de steekproeftrekking is beschreven en achteraf eenduidig reproduceerbaar, en ten minste 30% van alle patiënten uit de steekproef leveren volledige gegevens aan en er is geen statistische correctie uitgevoerd. *Of* De procedure voor de steekproeftrekking is beschreven en achteraf eenduidig reproduceerbaar, en minder dan 30% van alle patiënten uit de steekproef leveren volledige gegevens aan en er is een statistische correctie uitgevoerd.	
Voldoet niet	De procedure voor de steekproeftrekking laat ruimte voor selectie door de zorgaanbieder. *Of* Er wordt gebruik gemaakt van een meetperiode die korter is dan de periode waarover verslag wordt gedaan, en over alle patiënten worden gegevens aangeleverd. *Of* De procedure voor de steekproeftrekking is beschreven en achteraf eenduidig reproduceerbaar, en minder dan 30% van alle patiënten uit de steekproef leveren volledige gegevens aan en er is geen statistische correctie uitgevoerd.	Een indicator voor vochtletsel die in een - door de zorgaanbieder te kiezen - meetweek wordt vastgelegd en waarbij 'ernstig' zieke patiënten kunnen worden geëxcludeerd.

6. Statistisch betrouwbaar onderscheiden

6.1. Definitie

Statistisch betrouwbaar onderscheiden
Het vermogen van een indicator om zorgaanbieders met boven-gemiddelde en ondergemiddelde indicatorwaarden te onderscheiden van gemiddeld scorende aanbieders.

Een goede registratievergelijkbaarheid kan veel meetfouten voorkomen. Er zal echter altijd sprake zijn van toevallige verschillen die niet kunnen worden voorkomen door scherpere definities en strakkere registratieprocessen. Verschillen in indicatorwaarden tussen zorgaanbieders berusten hierdoor voor een deel op toeval. Toeval kan het gevolg zijn van toevallige variatie in de steekproef (sampling variability) en van toevalprocessen, anders dan zorgprocessen, die de uitkomst bij een patiënt mede bepalen (stochastic counterfactuals).

Een meting is *statistisch betrouwbaar* wanneer de indicatorwaarde niet gevoelig is voor toeval. Dat wil zeggen dat bij een herhaalde meting de indicatorwaarde maar weinig van de eerdere meting zal afwijken. Een zorgaanbieder kan zich *statistisch betrouwbaar onderscheiden* van het gemiddelde wanneer toeval de afwijking van de indicatorwaarde met het gemiddelde waarschijnlijk niet veroorzaakt (SiRM 2009).

6.2. Methoden om statische betrouwbaarheid te verbeteren

6.2.1. Wanneer speelt toeval een rol?

De invloed van een steekproef op de toevallige schatting zal kleiner zijn naarmate de grootte van de steekproef de grootte van de populatie nadert. Indien de volledige patiëntenpopulatie wordt gemeten dan speelt de toevallige steekproef geen rol meer. Daarmee is de invloed van toeval niet verdwenen. De uitkomst van een behandeling van een specifieke patiënt staat in veel gevallen niet van tevoren vast. Een groot aantal factoren kan een rol spelen en vele van die factoren hebben een toevalscomponent. Hierdoor blijft toeval en de bijbehorende statistiek een rol spelen, zelfs wanneer alle patiënten van een zorgaanbieder gedurende het gehele jaar zijn zijn gemeten.

Sommige indicatoren zijn echter niet gevoelig voor toevalsprocessen die verbonden zijn aan patiënten. Het gaat daarbij om indicatoren die los staan van patiëntkenmerken. Bij dergelijke indicatoren kan toeval wel een rol spelen bij het tot stand komen van kwaliteit van zorg, bijvoorbeeld het toevallige uitvallen van een verpleegkundige door een ski-ongeval. Vanuit het patiëntperspectief is het echter irrelevant hoe de kwaliteit tot stand is gekomen. De vraag is immers welke kwaliteit uiteindelijk is geboden.

Toeval speelt geen rol bij indicatoren waarbij (SiRM 2010b):

1. de patiëntkarakteristieken geen rol spelen, en
2. waarvan alle relevante geleverde zorg gedurende het jaar is gemeten.

Vanuit andere gebruikers dan de patiënt kan het onderscheid wat wel en wat niet als bron van toevalsvariatie wordt gezien verschillend zijn. Zie hiervoor bijlage 3.

Invloed van methoden op toeval
Het is belangrijk om al tijdens het ontwikkelen van een indicator na te denken over de statistische betrouwbaarheid of het statistische vermogen tot onderscheiden van een indicator. Het kan immers zo zijn dat het simpelweg niet mogelijk is om afdoende aantallen waarnemingen per zorgaanbieder te realiseren. Tegelijkertijd zijn er verschillende methodieken beschikbaar om de statistische betrouwbaarheid en het onderscheidingsvermogen van een indicator te verhogen.

Keuzes in registratie, rapportage-systematiek of data-aggregatie kunnen een grote impact hebben op de statistische betrouwbaarheid en het onderscheidingsvermogen van indicatoren. Daarbij gaat het om de gekozen uitkomstmaat, de daarbij behorende afkappunten, de steekproefomvang, het samenvoegen van metingen of meetmomenten, en het classificeren van kwaliteitsscores.
Naast deze keuzen speelt ook de statistische analysetechniek een belangrijke rol. Empirical Bayes is een methode die de factor 'toeval' in kwaliteitsmetingen, bij kleine proporties en weinig waarnemingen, beter kan filteren dan de frequentistische methoden.

6.2.2. Gekozen uitkomstmaat en afkappunten
Variabelen kunnen gemeten worden op een dichotome (ja/nee uitkomst), ordinale (uitkomsten kunnen geordend worden maar het is onduidelijk of verschillen in uitkomsten een vergelijkbare afstand hebben) of een intervalschaal (uitkomsten zijn geordend en verschillen hebben een vergelijkbare afstand). Door een dichotome schaal te gebruiken, gaat informatie verloren. In een ordinale of intervalschaal wordt de uitkomst in 'getal' geregistreerd, en is de spreiding in de uitkomsten vrijwel altijd groter. Informatieverlies vertaalt zich in een afname van het onderscheidingsvermogen.

6.2.3. Minimaal aantal waarnemingen (steekproefomvang)
Met behulp van een poweranalyse is het mogelijk om het minimaal aantal benodigde waarnemingen in een steekproef te bepalen. Een poweranalyse bepaalt bij hoeveel waarnemingen een zorgaanbieder met een betere (of slechtere) indicatorwaarde dan gemiddeld als significant verschillend wordt aangeduid.

Bij een poweranalyse is het *onderscheidingsvermogen* van een indicator van belang. Het *onderscheidingsvermogen* van een indicator hangt nauw samen met de statistische betrouwbaarheid. Waar de statistische betrouwbaarheid gaat over de kans dat een verschil door toeval als 'significant' wordt geduid, gaat het bij het onderscheidingsvermogen over de kans dat een werkelijk verschil door toeval *niet* als significant wordt geduid. Het onderscheidingsvermogen van een indicator (de 'power') is de kans dat een goed (of slecht) presterende zorgaanbieder te onderscheiden is van de gemiddeld presterende zorgaanbieders. In de wetenschap gaat men meestal uit van een onderscheidingsvermogen van 80%. Dat wil zeggen dat 20% van de boven- of ondergemiddeld presterende zorgaanbieders ten onrechte als 'gemiddeld' zullen worden aangemerkt.

Voor het waar te nemen verschil wordt de Cohen's effect size (ook wel Cohen's d) gehanteerd. Een voordeel van Cohen's d is dat er wetenschappelijke overeenstemming bestaat over wanneer een effect klein, medium of groot wordt genoemd. Voor de groene signaalvlag (goede statistische betrouwbaarheid) geldt een klein effect als criterium (Cohen's d = 0.2) en voor een gele signaalvlag (voldoet deels statistische betrouwbaarheid) een medium effect (Cohen's d = 0.5). Is ook een medium effect statistisch onvoldoende betrouwbaar vast te stellen, dan wordt een rode vlag toegekend. Doordat bij de berekening van de steekproefgrootte wordt uitgegaan van de Cohen's d, is de berekening niet gevoelig voor de richting van het effect.

Voor indicatoren die als proportie worden uitgedrukt, is het benodigde aantal observaties sterk afhankelijk van de proportie waarmee wordt vergeleken (in casu het landelijk gemiddelde). Om de complexiteit en het aantal rekenregels voor de bepaling van de signaalvlag beperkt te houden, is ervoor gekozen hier vijf categorieën te hanteren.

Een indicator scoort een groene vlag ('voldoet') indien het aantal waarnemingen groot genoeg is om een klein effect vast te stellen en een gele vlag ('voldoet deels') wanneer het aantal waarnemingen groot genoeg is om een medium effect vast te stellen.

Proportie De gemiddelde score op de indicator (teller gedeeld door noemer)		**Statistische betrouwbaarheid**	**Benodigde observaties** (noemer)
P≤0,02	p≥0,98	Voldoet (groen)	≥ 2300
		Voldoet deels (geel)	Tussen 640 en 2300
		Voldoet niet (rood)	< 640
0,02<p≤0,04	0,98>p≥0,96	Voldoet (groen)	≥ 1160
		Voldoet deels (geel)	Tussen 320 en 1160
		Voldoet niet (rood)	< 320
0,04<p≤0,07	0,96>p≥0,93	Voldoet (groen)	≥ 580
		Voldoet deels (geel)	Tussen 160 en 580
		Voldoet niet (rood)	< 160
0,07<p≤0,12	0,93>p≥0,88	Voldoet (groen)	≥ 340
		Voldoet deels (geel)	Tussen 90 en 340
		Voldoet niet (rood)	< 90
0,12<p≤p<0,88		Voldoet (groen)	≥ 200
		Voldoet deels (geel)	Tussen 30 en 200
		Voldoet niet (rood)	< 30

Tabel 3 Overzicht van benodigd aantal waarnemingen om proporties betrouwbaar te onderscheiden

Tabel is overgenomen uit Onderzoeksverantwoording, Statistisch betrouwbaar onderscheiden, Significant 2011.

Uitgaande van een landelijk gemiddeld decubitusniveau van 3% dient een zorgaanbieder conform deze methodiek over minimaal 1160 waarnemingen te beschikken om 'goed' betrouwbaar onderscheidend te zijn, en over ten minste 320 waarnemingen om 'matig' betrouwbaar onderscheidend te zijn.

6.2.4. Samenvoegingen

Door indicatoren per zorgaanbieder samen te voegen kan het vermogen om statistisch betrouwbaar te onderscheiden toenemen. Dit werkt echter alleen wanneer de indicatoren van een zorgaanbieder op de gemeten onderwerpen sterk positief met elkaar samenhangen. In dat geval is een positieve gemiddelde score een sterke voorspeller voor een positieve score op een enkele indicator. Hoe hoger de samenhang is, hoe meer zorgaanbieders bij een meting statistisch betrouwbaar kunnen worden onderscheiden. Als indicatoren niet of nauwelijks positief samenhangen, worden de indicatoren waarop slecht wordt gescoord, gecompenseerd door indicatoren waarop goed wordt gescoord. In dat geval neemt het statistisch betrouwbaar onderscheiden weinig toe en mogelijk zelfs af.

Het is daarom aantrekkelijker om indicatoren samen te voegen die zowel inhoudelijk als statistisch samenhangen. Zo is het te verwachten dat een goede hygiëne in een ziekenhuis leidt tot minder infecties bij verschillende behandelingen, of kan worden verwacht dat een tekort aan personeel in een verpleeghuis leidt tot meer fixatie en meer decubitus[15].

[15] Indien men kiest voor samengestelde indicatoren is het mogelijk om losse indicatoren meerdere malen te gebruiken voor verschillende samengestelde indicatoren. Op deze wijze kan efficiënt gebruik worden gemaakt van de verzamelde indicatoren en kan een relatief breed palet aan samengestelde indicatoren aan de vereiste voor statistisch betrouwbaar onderscheiden voldoen. Wanneer in de praktijk blijkt dat deze samengestelde

In de internationale literatuur is nog weinig bekend over het gebruik van composietindicatoren in de gezondheidszorg en de daarbij behorende methodologische eisen. Het is wel mogelijk met de data schaalconstructies te maken (met factoranalyse en *concept mapping*). Bij de uitwerking van de samengestelde indicatoren moeten in ieder geval de volgende punten worden uitgewerkt:

- beschrijving van wat een samengestelde indicator is en welk gebruikersdoel dit dient;
- de methodologische eisen waarmee tijdens de ontwikkeling rekening moet worden gehouden;
- hoe een samengestelde indicator moet worden beoordeeld in termen van validiteit, betrouwbaarheid en populatievergelijkbaarheid etc.

6.2.5. Gebruik van voor- en nametingen

Toevalsvariatie hangt in veel gevallen sterk samen met (niet geobserveerde) eigenschappen van de patiënt. Deze variatie kan uit de meting worden gefilterd met behulp van een voor- en nameting. Door de resultaten van de voormeting van de nameting af te trekken ontstaat een verschil. Dat verschil is in de praktijk veelal minder gevoelig voor toevalsvariatie en levert daarom betrouwbaardere metingen op. De voormeting is bij voorkeur een nulmeting bij de patiënt voorafgaand aan de behandeling. Een nameting vindt bij voorkeur plaats op het moment dat een goed beeld ontstaat van het effect van de behandeling. Ook kan het gaan om meerdere metingen bij dezelfde patiënt gedurende een bepaalde periode.

6.2.6. Omgaan met kleinere aantallen waarnemingen: empirical Bayes[16]

Het komt regelmatig voor dat informatie over de kwaliteit van de geleverde zorg gevraagd wordt over kleinere (afdelingen binnen) zorgaanbieders, aandoeningen die relatief weinig voorkomen en/of weinig voorkomende risico's. Voorbeelden hiervan zijn het meten van (gewelds)incidenten in de gehandicaptenzorg, uitkomsten bij de zorg voor Cystic Fibrose (taaislijmziekte), of de decubitus scores in kleinschalige woonarrangementen.

Juist vanwege de rol die toeval speelt bij kleine aantallen waarnemingen wordt in onderzoek naar kwaliteit van zorgaanbieders in de regel gebruik gemaakt van empirical Bayes technieken. Dit wordt gezien als alternatief voor de frequentistische technieken die de basis vormen van de hierboven beschreven invulling van statistische betrouwbaarheid. De veronderstelling achter de empirical Bayes technieken is dat de kwaliteit van zorgaanbieders met elkaar samenhangt. Met andere woorden, wanneer alle gemeten verpleeghuizen een decubitus incidentie hebben tussen de 0 en de 7%, dan is het waarschijnlijk dat de incidentie van een verpleeghuis dat nog geen gegevens heeft aangeleverd, ook tussen de 0 en de 7% ligt. Bij de schatting van de incidentie van het verpleeghuis dat nog geen gegevens heeft aangeleverd wordt gebruik gemaakt van de gegevens van de al gemeten verpleeghuizen. Stel dat een verpleeghuis maar één patiënt onderzoekt en deze blijkt decubitus te hebben, dan is de frequentistische schatting voor deze zorgaanbieder 100%. Wordt er gebruik gemaakt van de empirical Bayes methode, dan zal deze schatting niet 100% zijn, maar richting de 7% worden getrokken.

Met empirical Bayes technieken maakt men schattingen op basis van de waarnemingen bij een zorgaanbieder zelf én de verdeling van waarnemingen bij alle andere zorgaanbieders. Naarmate meer waarnemingen bij een zorgaanbieder zijn gedaan, wordt het relatieve belang van die waarnemingen groter. Het belangrijkste voordeel van deze techniek is de vermindering van de invloed van toeval op de indicatorwaarde. Anders gezegd: het gebruik van empirical Bayes technieken leidt ertoe dat afwijkingen, die vooral het gevolg zijn van toeval, krimpen naar het gemiddelde (*shrinkage)*. Dit krimpen zal ertoe leiden dat zorgaanbieders met (heel) weinig

---

indicatoren met elkaar correleren dan leidt samenvoeging tot een toenamen van het statistisch betrouwbaar onderscheiden van een interpreteerbare maat. Een voorbeeld van een samengestelde indicator is de CQ-index methodiek waarvoor o.a. Likert-schalen worden geconstrueerd op basis van met elkaar samenhangende items.

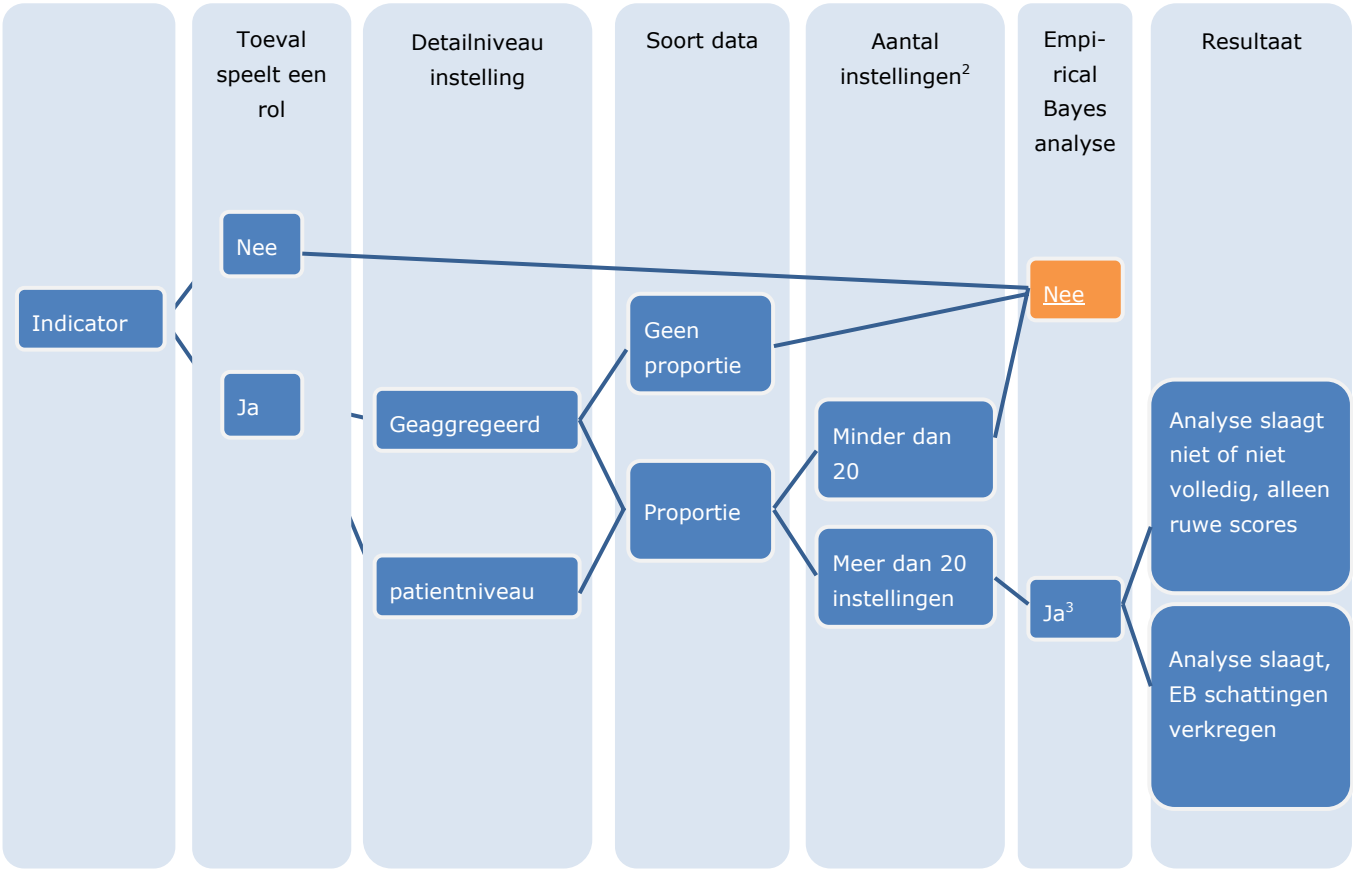
[16] Deze paragraaf is gebaseerd op Koolman et al. (2010), SiRM (2010a), de Brouwer et al. (2010) en Significant (2011).

waarnemingen minder snel extreme kwaliteitsoordelen krijgen, en daarmee dat (kleine) zorgaanbieders van jaar tot jaar sterk uiteenlopende scores vertonen.

De empirical Bayes-methodiek is inmiddels in verschillende zorgsectoren getest, en behoort inmiddels ook voor de CQ-index methodiek tot de standaard methodieken. Empirical Bayes-technieken dienen daarom tot de standaardmethodiek van het analyseren van indicatorwaarden te behoren. Toch kan empirical Bayes niet altijd worden toegepast.

Er is een beslismodel opgesteld wanneer empirical Bayes kan worden toegepast (Significant, 2011). Het model wordt weergegeven in figuur 6. De beslissing is afhankelijk van het niveau waarop de data wordt verzameld (individueel of geaggregeerd), het soort data (proporties of anders), het aantal instellingen waarvan de kwaliteit wordt gemeten (groter of kleiner dan 20) en het al of niet slagen van de analyse.

figuur 6 Beslismodel empirical Bayes



1 Dit beslismodel is ontworpen voor indicatoren gemeten op een dichotoom of interval schaal.
2 Het streven blijft om bij minimaal 30 zorgaanbieders metingen te verrichten en minimaal 30 observaties per zorgaanbieder.
3 Bij proporties dient uit te worden gegaan van een normaal-binomiaal model, bij gemiddelden van een normaal-normaal model, en bij beide optimalisatie met het Newton-Rahpson algoritme

6.3. Statistische betrouwbaarheid in relatie tot het type indicator

Voor *structuurindicatoren* speelt toeval per definitie geen rol, omdat de te leveren informatie niet gebaseerd is op een serie van afzonderlijke waarnemingen bij patiënten, maar over een waarneming van een structuurkenmerk van de aanbieder.

Voor *procesindicatoren* speelt toeval een rol daar waar de indicatorwaarden van een procesindicator (over bijvoorbeeld doorlooptijden) kunnen afhangen van toevallige verschillen in samenstelling van de patiëntenpopulatie. In sommige gevallen, echter, is een procesindicator niet of nauwelijks afhankelijk van de patiëntenkenmerken, maar primair van de kenmerken van de organisatie zelf. Een voorbeeld hiervan is de 'door-to-needle time' bij de acute behandeling van CVA met thrombolytica voor geïndiceerde patiënten. In een dergelijk geval kan ook met een beperkt aantal waarnemingen een statistisch betrouwbare uitspraak over de geleverde kwaliteit worden gedaan.

Uitkomstindicatoren zullen in de praktijk altijd afhankelijk zijn van toeval, omdat toevalsprocessen de uitkomst bij een patiënt mede bepalen, ongeacht of alle patiënten gedurende het gehele jaar zijn betrokken bij de kwaliteitsmeting.

6.4. Beoordelen statistisch betrouwbaar onderscheiden

In de onderstaande tabel staat beschreven, en met voorbeelden geïllustreerd, hoe de signaalvlaggen groen, geel en rood voor het criterium statistisch betrouwbaar onderscheiden toegekend kunnen worden. Dit criterium van de beoordeling wordt per indicatorset en per indicator beoordeeld. Er wordt dus onderscheid gemaakt naar indicatoren. Bij de beoordeling wordt uitgegaan van het vergelijken van de indicatorwaarde van een zorgaanbieder met de gemiddelde verwachte waarde.

Vlag	Toelichting	Voorbeeld
Voldoet	Het is aangetoond (ex-post) of onderbouwd (ex-ante) dat voor 80% van de boven of onder gemiddeld presterende zorgaanbieders met een 'kleine afwijking' en met 5% significantieniveau geconcludeerd kan worden dat deze boven- of ondergemiddelde score niet op toeval berust.	Uit de IGZ-registraties blijkt dat het landelijk gemiddelde percentage postoperatieve patiënten met op enig moment ernstige pijn op 5,6% ligt (2008)[17]. Een gemiddelde zorgaanbieder heeft volgens deze gegevens dan bij 2.780 postoperatieve patiënten de pijn gemeten (excl. dagbehandeling)[31]. Voor aanbieders met meer dan 580 waarnemingen kan worden gesteld dat ze voldoende waarnemningen hebben om een klein effect (afwijking) significant te observeren.
Voldoet deels	Het is aangetoond (ex-post) of onderbouwd (ex-ante) dat voor 80% van de boven of onder gemiddeld presterende zorgaanbieders met een medium afwijking en 5% significantieniveau geconcludeerd kan worden dat deze boven- of ondergemiddelde score niet op toeval berust.	Voorbeeld hierboven maar dan voor zorgaanbieders met tussen de 180 en 560 waarnemingen.
Voldoet niet	Statistisch betrouwbaar onderscheiden is niet afdoende onderbouwd, en het aantal waarneming is onvoldoende om	Voorbeeld hierboven maar dan voor zorgaanbieders met minder dan 180 waarnemingen.

[17] Het resultaat telt 2008! Kwaliteitsindicatoren als onafhankelijke graadmeter voor de kwaliteit van in ziekenhuizen verleende zorg. IGZ 2009

een medium sterke afwijking
van het verwachtte gemiddelde
vast te stellen.

De beoordeling van de statistische betrouwbaarheid van een indicator - voor alle zorgaanbieders gezamenlijk - wordt gebaseerd op de signaalvlaggen per zorgaanbieder:

- Wanneer 25% of meer van de zorgaanbieders een rode signaalvlag heeft, dan is een zinnige vergelijking van de zorgaanbieders met het landelijk gemiddelde niet mogelijk en krijgt deze indicator een rode signaalvlag;
- Indien 75% of meer van de zorgaanbieders een groene signaalvlag heeft, dan krijgt de indicator een groene signaalvlag en is een zinnige vergelijking van de zorgaanbieders met het landelijk gemiddelde mogelijk;
- Als de signaalvlag voor de indicator niet rood en niet groen is. Dus: minder dan 25% van de zorgaanbieders heeft een rode signaalvlag en minder dan 75% van de zorgaanbieders heeft een groene signaalvlag, dan krijgt de indicator een gele signaalvlag en vraagt een vergelijking van de zorgaanbieders om een genuanceerd oordeel.

Literatuur

AHRQ. 'Refinement of the HCUP Quality Indicators'. http://www.ahrq.gov/clinic/tp/hcupqitp.htm, geraadpleegd op 1 februari 2012.

Beersen, N., Kallewaard M., van Croonenborg J.J., J.J.E. van Everdingen, T.A. van Barneveld. (2007) Handleiding Indicatorontwikkeling, Den Haag: ZonMw.

Berg, M., Ed. (2004) Health Information Management: Integrating Information and Communication Technology in Health Care Work. London, Routledge.

Berg, M., Klazinga N., et al. (2001) Van 'evidence-based' naar 'value-based': Normatieve overwegingen bij richtlijnen voor Passende Medische Zorg. In Ingebouwde Normen. Medische Technieken Doorgelicht. M. Berg and A. Mol. Utrecht, Van der Weest.

De Boer D., Van der Hoek L., Delnoij D., Groenewegen P. (2010) Kleine zorgaanbieders in multilevel vergelijkende analyses. De CQI Verpleging, Verzorging en Thuiszorg. Utrecht: Nivel.

Bosch W van den, 2011, De HSMR beproefd: Aard en invloed van meetfouten bij het bepalen van het gestandaardiseerde ziekenhuissterftecijfer, Proefschrift, Vrije Universiteit.

Campbell, S. M., Braspenning, J., et al. (2002) Research methods used in developing and applying quality indicators in primary care. Qual Saf Health Care **11**(4): 358-64.

Kwaliteitsinstituut voor de gezondheidszorg CBO (2005) Evidence-based richtlijn ontwikkeling: handleiding, Utrecht. (www.cbo.nl/thema/Richtlijnen/EBRO-handleiding).

Committee on Redesigning Health Insurance Performance Measures, P., and Performance Improvement Programs, (2006) Performance measurement: accelerating improvement, Washington DC, National Academies Press.

Centrum Klantervaring Zorg, Handboek CQ Ontwikkeling en Eisen en Werkwijzen Metingen, http://www.centrumklantervaringzorg.nl/cqi-richtlijnen/handboek-eisen-en-werkwijzen-cqi-metingen.html. geraadpleegd op 1 februari 2012.

Fink A., Kosecoff J., Chassin M., Brook R.H. (1984) Consensus Methods: Characteristics and Guidelines for Use. American Journal of Public Health. 74(9):p. 979-983.

Iezzoni, L. I., Ed. (2003) Risk Adjustment for Measuring Health Care Outcomes. Chicago, Health Administration Press.

Hernan MA, Robins JM (2006) "Estimating Causal Effects From Epidemiological Data". Journal of Epidemiology & Community Health, 60;578-596).

Jones J.J., Hunter D. (1995) Consensus methods for medical and health services research. British Medical Journal 311:p. 376–80.

Kallewaard, M., Beersen N., van Everdingen J.J.E., van Croonenborg J.J., van Barneveld T.A. (2007) Kwaliteit van zorg in de etalage: eindrapport. Den Haag: ZonMw.

De Koning J., Smulders, A., Klazinga N.S. (2007) Appraisal of indicators through Research and Evaluation (AIRE) 2.0, 2007. Amsterdam: Academisch Medisch Centrum Universiteit van Amsterdam, afdeling sociale geneeskunde.

Koolman X., Luijendijk H., Boonen L. (2011) Op weg naar meer betrouwbare prestatiemeting in verpleeghuizen, verzorgingshuizen en thuiszorgorganisaties. Tijdschrift voor Ouderengeneeskunde, 2: 47-54.

Krumholz, H.M., et al. (2007) Measuring Performance For Treating Heart Attacks And Heart Failure: The Case For Outcomes Measurement. Health Affairs, 26(1): p. 75-85.

Lu, M., Ma C.T. (2002) Consistency in performance evaluation reports and medical records, Journal of Mental Health Policy and Economics, 2002. 5(4):141-52.

Mant J. (2001) Process versus outcome indicators in the assessment of quality of health care, International Journal for Quality in Health Care, 13(6): 475-480.

NHS information centre, 'Patient Reported Outcomes Measures (PROMs)' http://www.ic.nhs.uk/proms, geraadpleegd op 1 november 2010.

Nicholas L.H., N.H. Osborne, MD; J.D. Birkmeyer, Justin B. Dimick, (2010) Hospital Process Compliance and Surgical Outcomes in Medicare Beneficiaries, Archives of Surgery, 145(10): 999-1004

Rothman KJ, S Greenland, TL Lash (2008) Modern Epidemiology, Lippincott Williams & Wilkins; Third edition.

Significant (2011) Statistisch betrouwbaar onderscheiden: Onderzoeksverantwoording. Studie uitgevoerd door Significant in opdracht van Zichtbare Zorg.

SiRM, (2008) Standaardisatie zorginhoudelijke indicatoren verpleging, verzorging en zorg thuis. Studie uitgevoerd door SiRM in opdracht van Zichtbare Zorg.

SiRM (2010a) Standaardisatie zorginhoudelijke indicatoren verpleging, verzorging en zorg thuis. 3e Meetronde 2009. Studie uitgevoerd door SiRM in opdracht van Zichtbare Zorg.

SiRM (2010b) Beschrijving toetsingskader standaardisatie en presentatie van de zorginhoudelijke indicatoren GZ. Studie uitgevoerd vdoor SiRM in opdracht van Zichtbare Zorg.

VWS (2011) Inhoudelijke kaderstelling voor het transparantieprogramma, Ministerie van VWS.

Rachel M. Werner, Eric T. Bradlow (2006) Relationship Between Medicare's Hospital Compare Performance Measures and Mortality Rates, Journal of the American Medical Association, 296(22):2694-2702.

Zichtbare Zorg (2008) Voorkomen is beter dan genezen. Betrouwbaarheid van kwaliteitsinformatie in de zorg: achtergrondstudie naar risico's en oplossingsrichtingen. Studie uitgevoerd door PwC/TNO in opdracht van het programma Zichtbare Zorg, Zichtbare Zorg, Den Haag

Zichtbare Zorg (2009a) Vergroten vergelijkbaarheid data prestatie-indicatoren GGZ (Eindrapportage) Studie uitgevoerd door Plexus in opdracht van Zichtbare Zorg, Zichtbare Zorg, Den Haag.

Zichtbare Zorg (2009b) Ontwikkeling van een methodiek om de kwaliteit van aangeleverde data te beoordelen. Zichtbare Zorg, Den Haag.

Zichtbare Zorg (2009c) Betrouwbaar onderscheiden. Een achtergrondstudie naar de statistische betrouwbaarheid en steekproefomvang bij het vergelijken van zorgaanbieders. Studie uitgevoerd door X. Koolman in opdracht van Zichtbare Zorg. Zichtbare Zorg, Den Haag.

Zichtbare Zorg (2009d) Het Raamwerk Kwaliteitsindicatoren. Zichtbare Zorg, Den Haag.

Zichtbare Zorg (2009e) Het ordeningskader Ziekenhuis Indicatoren. Zichtbare Zorg, Den Haag.

Zichtbare Zorg (2009f) Vergroten vergelijkbaarheid data prestatie-indicatoren GGZ, Studie uitgevoerd door Plexus in opdracht van Zichtbare Zorg, Den Haag.

Bijlage 1: Dankzegging

Voor de totstandkoming van de herziening van de indicatorstandaard 1.5 en 2.0 gaat speciale dank uit naar:

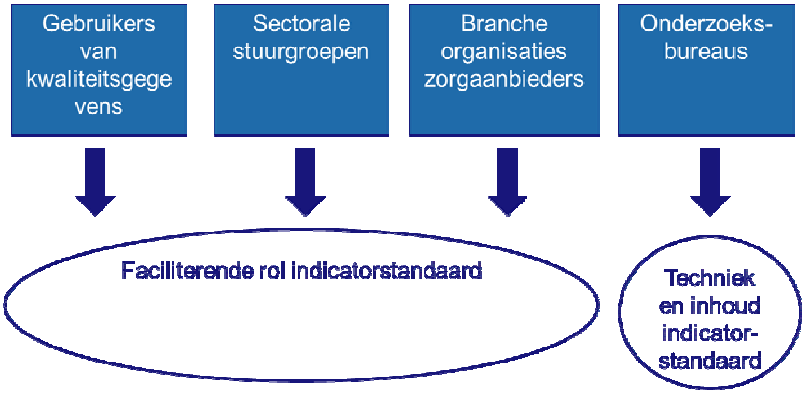
Marlies Bennema	Orde van Medisch Specialisten
Marc Berg	KPMG Plexus
Loes Bierma	Nederlandse Patiënten Consumenten Federatie
Margriet Bouma	Nederlands Huisartsen Genootschap
Anemone Bögels	Nederlandse Federatie van Kankerpatiëntenorganisaties
José Braspenning	IQ healthcare
Diana Delnoij	Centrum Klantervaring Zorg
Marie Jose Driessen	Vereniging Gehandicaptenzorg Nederland
Loes Koster	Significant
Lisenka van Loon	Nederlandse Federatie van Kankerpatiëntenorganisaties
René Oude Vrielink	NVZ vereniging van ziekenhuizen
Maria Schipper	Significant
Jeroen Schols	Revalidatie Nederland
Jaap Schrieke	GGZ Nederland
Andy Schuurmans	Nederlandse Federatie van Universitair Medische Centra
Sjoerd Terpstra	Zorgverzekeraars Nederland
Martine Versluijs	Nederlandse Patiënten Consumenten Federatie
Marion Verduijn	NFU -consortium 'Kwaliteit van Zorg'

en de leden van de Statistiek Commissie van het Centrum Klantervaring Zorg.

Bijlage 2: Proces evaluatie indicatorstandaard 1.0

De indicatorstandaard 1.0 is aangepast naar een versie 1.5. Daarvoor heeft eerst een evaluatie van de indicatorstandaard 1.0 plaatsgevonden. De evaluatie heeft in stappen plaatsgevonden. Het doel van de evaluatie was enerzijds de faciliterende rol van de indicatorstandaard te verbeteren, anderzijds om de techniek van de standaard aan te passen aan de huidige stand van de wetenschap en de bevindingen van de ontwikkeling van kwaliteitsindicatoren in sectoren. Hiervoor is onderscheid gemaakt naar vier verschillende groepen van belanghebbenden:

1. beoogde gebruikers van de indicatorwaarden (patiëntenorganisaties, zorgverzekeraars);
2. sectorale stuurgroepen van Zichtbare Zorg;
3. brancheorganisaties van zorgaanbieders;
4. onderzoeksbureaus.



Met de beoogde gebruikers van de indicatorwaarden zijn in totaal drie partijen geïnterviewd: ZN, (Sjoerd Terpstra) NPCF (Loes Bierma en Martine Versluijs) en NFK (Anemone Bögels en Lisenka van Loon). De kernvraag van de interviews met ZN, NPCF en NFK was boven water te krijgen op welke manier de indicatorstandaard beter kan aansluiten op de wensen van de geïnterviewde partij. Tijdens de interviews bleek dat de indicatorstandaard niet goed bekend was bij de geïnterviewden. SiRM heeft daardoor tijdens de interviews ook een toelichting gegeven op de indicatorstandaard.

De sectorale stuurgroepen zijn individueel via de projectleiders van Zichtbare Zorg benaderd. Afhankelijk van de stand van zaken in verschillende stuurgroepen is de indicatorstandaard besproken. Voor twee sectoren heeft hiervoor een apart gesprek plaatsgevonden.

Met een aantal brancheorganisaties van zorgaanbieders hebben twee discussiebijeenkomsten plaatsgevonden. De eerste discussiebijeenkomst stond in het teken van de vraag "Op welke manier kan de indicatorstandaard beter aansluiten bij de wensen van de zorgaanbieders?". In de tweede discussiebijeenkomst heeft SiRM een terugkoppeling gegeven van de wijze waarop de input is verwerkt. De input heeft met name effect gehad op de routekaart. Naar aanleiding van de tweede discussiebijeenkomst heeft SiRM de kaart verder kunnen aanscherpen. Naast deze bijeenkomsten hebben twee branchepartijen ook een schriftelijke reactie op de indicatorstandaard 1.0 gestuurd. De discussiebijeenkomsten kenden de volgende deelnemers: Margriet Bouma (NHG), Maartje Blom (VGN), Jaap Schrieke (GGZ Nederland), Marlies Bennema (Orde van Medisch Specialisten, René Oude Vrielink (NVZ) en Andy Schuurmans (NFU).

Met 4 onderzoeksbureaus/onderzoeksinstituten (KPMG Plexus: Marc Berg, Significant: Maria Schipper & Loes Koster, IQ healthcare: José Braspenning en het CKZ: Diana Delnoij) zijn interviews gehouden. De kernvraag tijdens de interviews was "Wat zijn de bevindingen met de indicatorstandaard 1.0 en hoe kan de indicatorstandaard verbeterd worden?"

Naast de genoemde gesprekken is twee keer met het Ministerie van VWS (Kees Molenaar, Ivana Gomes-Durao, Cynthia Vogeler) gesproken. Deze gesprekken gingen enerzijds over de doelstelling van het Ministerie van VWS en anderzijds over de inbedding van de indicatorstandaard.

De interviews en de discussiebijeenkomsten hebben veel informatie voor SiRM opgeleverd die verwerkt is in *Routekaart* en *Indicatorstandaard*. De onderzoekers danken de geïnterviewden en de deelnemers aan de discussiebijeenkomsten voor hun actieve en constructieve bijdrage aan de doorontwikkeling. Daarnaast hebben de onderzoekers veel baat gehad van de suggesties van het team Kennis en Kwaliteit van Zichtbare Zorg (Eline Meijer, Femke Vlems, Elske Faber).

Bijlage 3: Overige gebruiksdoelen van indicatorwaarden

De indicatorstandaard beschrijft een toetsingskader voor de evaluatie van kwaliteitsindicatoren vanuit een patiëntkeuzeperspectief. De indicatoren dienen de patiënt, zijn verzorger of verwijzer te ondersteunen bij het kiezen van een zorgaanbieder. De keuze voor dit gebruiksdoel staat het gebruik van de data voor andere doelen niet in de weg. Wel kan het voor de overige gebruikers nuttig zijn om te weten welke keuzen zouden zijn gemaakt wanneer de indicatoren een ander doel zouden dienen. Het onderstaande overzicht is niet uitputtend maar geeft inzichten in de verschillen tussen gebruiksdoelen. Het ovezicht geeft aan in welke mate kwaliteitsindicatoren vanuit patiënt-keuzeperspectief voor andere gebruikersdoelen aan de beoordelingscriteria voldoet. De volgende gebruiksdoelen worden onderscheiden:

I. selectieve en prestatie-contractering;
II. zorgverbetering;
III. toezicht.

I Selectieve en prestatie-contractering

Zorgverzekeraars kunnen indicatoren gebruiken voor het contracteren van zorg. Het is waarschijnlijk dat de voorkeuren van patiënten een belangrijke rol spelen bij het contracteren van zorg. Valide en betrouwbare patiëntkeuze-indicatoren kunnen daarom gebruikt worden bij het selecteren van zorgaanbieders en bij het opstellen van prestatiecontracten. Voor de contractering is het echter waarschijnlijk dat meerdere indicatoren gezamenlijk gebruikt worden om tot een oordeel te komen. Een dergelijk samengesteld valide en betrouwbaar oordeel kan gebaseerd worden op een mix van meer en minder valide en betrouwbare indicatoren.

Inhoudsvaliditeit

De inhoudsvaliditeit van een indicator(set) hangt samen met de kwaliteitsvisie. Zorgverzekeraars Nederland heeft in Visie op Kwaliteit[18] aangegeven dat kwaliteit van zorg staat voor:

1. medisch inhoudelijke kwaliteit;
2. patiëntgerichtheid;
3. doelmatigheid.

De eerste twee pijlers komen overeen met de visie op kwaliteit zoals die is verwoord in de meest visiedocumenten van de stuurgroepen. De laatste pijler wordt in de meest visiedocumenten niet genoemd, en maken waarschijnlijk geen deel uit van de beoordeling van de inhoudsvaliditeit. Doelmatigheid heeft betrekking op de kwaliteit/prijsverhouding van zorg.

Vertekening: populatievergelijkbaarheid

Voor selectieve contractering dienen zorgaanbieders onderling vergelijkbaar te zijn. Daarbij kan gebruik worden gemaakt van de indicatoren die gecorrigeerd zijn voor populatiekenmerken. Naast populatieverschillen zijn er echter nog andere factoren die goede of slechte uitkomsten bepalen en die voor de zorgaanbieder niet beïnvloedbaar zijn. Een voorbeeld hiervan is de kwaliteit van de lokale arbeidsmarkt. Tijdens de contractering kunnen deze factoren wel van belang zijn.

Prestatie-contractering kan gebaseerd zijn op (1) onderlinge vergelijking van de kwaliteit van zorg tussen zorgaanbieders of (2) een verbetering van de kwaliteit van zorg van een zorgaanbieder. Bij het belonen van een verbetering wordt de huidige kwaliteit vergeleken met eerdere geleverde kwaliteit. Daarbij is het niet nodig om te corrigeren voor populatieverschillen tussen

18

zorgaanbieders. Ongecorrigeerde indicatoren zijn eenvoudiger direct te koppelen aan geleverde zorg. Daardoor zijn ze aantrekkelijker voor zorgverbetercontracten.

Het gebruik van regressie als standaardisatietechniek sluit niet optimaal aan bij de informatievraag voor selectief inkopen. Het is echter nog onvoldoende duidelijk hoe groot de invloed van deze methodische keuze is op de bruikbaarheid van de gestandaardiseerde informatie voor selectieve zorginkoop.

Vertekening: registratievergelijkbaarheid

Registratievergelijkbaarheid is waarschijnlijk een probleem voor zowel afzonderlijke indicatoren als voor oordelen die gebaseerd zijn op meerdere indicatoren (samengestelde indicatoren). Het oordeel dat volgt uit de indicatorstandaard is daardoor relevant voor contractering. Bij een negatief oordeel kan aanvullend onderzoek wenselijk zijn. Een contracterende partij kan dit doen op basis van declaratiedata, of op basis van administratieve gegevens van de zorgaanbieders zelf.

Vertekening: steekproefvergelijkbaarheid

De steekproefvergelijkbaarheid is ook voor de contracterende partij relevant. Maar ook hier hebben de contractpartijen veelal aanvullende mogelijkheden om per zorgaanbieder te controleren of de resultaten representatief zijn door gebruik te maken van declaratiedata of van administratieve gegevens van de zorgaanbieders zelf.

Statistisch betrouwbaar onderscheiden

Voor een contract zal een afzonderlijke indicator waarschijnlijk niet relevant zijn; alle indicatoren die de kwaliteit van een bepaald product beschrijven zullen bij de contratering worden betrokken. Impliciet of expliciet wordt dan gebruik gemaakt van een samengestelde indicator. De relevante statistische vraag is dan of deze samengestelde indicator statistisch betrouwbaar onderscheidend is. Daarbij is het heel goed mogelijk dat de samengestelde indicator goed scoort op dit criterium terwijl de onderliggende indicatoren matig of slecht scoren.

Voor het berekenen van het statistisch betrouwbaar onderscheidend vermogen is bij voorkeur de ruwe data nodig (met maatregelen om de privacy van patiënten te borgen). Echter met een aantal aannamen kan ook worden gewerkt met geaggregeerde data. Indien empirical Bayes is gebruikt voor de afzonderlijke indicatoren dan wordt een dergelijke bewerking wezenlijk lastiger.

II Zorgverbetering

Zorgaanbieders, patiëntenraden en zorginkopers kunnen indicatoren gebruiken met het doel om de zorg te verbeteren. Vergelijkingen van de kwaliteit van een zorgaanbieder over de tijd vallen buiten het doel van de indicatorstandaard. De gegevens die verzameld worden voor de indicatoren zijn echter wel te gebruiken voor vergelijkingen over de tijd. Hieronder wordt de informatiebehoefte beschreven die nodig is om de kwaliteit van zorg te verbeteren.

Inhoudsvaliditeit

Uitkomstindicatoren leveren goede patiëntkeuze-informatie, maar geven geen inzicht in de kwaliteit van de onderliggende processen en hoe de zorg kan worden verbeterd. Daarom kunnen uitkomstindicatoren die inhoudsvalide zijn op individueel en op setniveau, vanuit een verbeterperspectief toch onvoldoende volledig zijn. Naast aanvullende procesindicatoren kan een goed registratiesysteem helpen bij het vertalen van uitkomsten naar verbeteracties. Met name registratiesystemen die de zorgaanbieder in staat stellen slechte uitkomsten te herleiden tot individuele patiënten en de geleverde zorg, kunnen het verbeterproces ondersteunen.

Vertekening: populatievergelijkbaarheid

De samenstelling van de patiëntpopulatie van een zorgaanbieder verandert in de regel slechts weinig over de tijd. Dat is een gevolg van de specialisatie en daarmee de positie in de markt, de

regio waaruit de patiënten komen, de samenwerking met andere zorgaanbieders in de omgeving en het verwijsgedrag van de verwijzende zorgaanbieders. Voor de vergelijkbaarheid van de uitkomsten tussen verschillende meetmomenten is een correctie voor populatieverschillen daarom veelal niet nodig. De vergelijking die vanuit het patiëntkeuzeperspectief wordt gemaakt is erop gericht zorgaanbieders op enig moment met elkaar te vergelijken. Deze correctie kan de vergelijkbaarheid over de jaren verslechteren in plaats van verbeteren. Daarnaast leidt de correctie voor minder transparante gegevens. Uitkomsten kunnen daardoor niet meer herleid worden naar patiënten en hun uitkomsten en de bruikbaarheid van informatie neemt hierdoor af. Indien de gegevens met empirical Bayes worden geanalyseerd dan zijn vergelijkingen tussen de tijd vrijwel onmogelijk. Voor zorgverbetering dient daarom vooral gebruik te worden gemaakt van ruwe (niet gecorrigeerde) gemiddelden om de ontwikkeling van kwaliteit van zorg te analyseren en te monitoren.

Vertekening: registratievergelijkbaarheid

Registratieverschillen hinderen de onderlinge vergelijking van zorgaanbieders. Echter, in veel gevallen zal binnen een instelling wel op een vergelijkbare wijze worden geregistreerd. Hierdoor zijn vergelijkingen tussen jaren of tussen behandelaren, ondanks registratieverschillen tussen zorgaanbieders, wel mogelijk. Daarnaast zijn zorgaanbieders in vele gevallen in staat om de kwaliteit van de eigen registratie op waarde te schatten. Een kritisch deskundigenoordeel over de registratievergelijkbaarheid hoeft het gebruik van een indicator voor zorgverbetering daarom niet in de weg te staan.

Vertekening: steekproefvergelijkbaarheid

Ook voor de steekproefvergelijkbaarheid geldt dat een kritisch deskundigenoordeel niet direct leidt tot de conclusie dat de gegevens onbruibaar zijn voor zorgverbetering. Zo is het mogelijk dat de oorzaken voor een niet representatieve steekproef gelegen zijn in het afwijkend hanteren van exclusiecriteria. Indien deze criteria elk jaar op dezelfde wijze worden gehanteerd dan zijn resultaten niet representatief voor de gehele behandelde populatie, maar wel bruikbaar om de kwaliteit van zorg te monitoren die geleverd is aan de waargenomen groep.

Statistisch betrouwbaar onderscheiden

Ook voor het monitoren van de kwaliteitsontwikkeling speelt toeval een rol. Zorgaanbieders beschikken echter over veel informatie waarmee zij een trend kunnen beoordelen. Zorgaanbieders kunnen in veel gevallen daardoor toch indicatorwaarden gebruiken waarover deskundigen het onderscheidingsvermogen bekritiseren. Een voorbeeld: Indien de indicatoren van een bepaalde afdeling tussen twee jaren sterk verslechteren dan kan dat het gevolg zijn van toeval. Als in het laatste jaar op die afdeling tevens sprake was van personele problemen dan is de kans groot dat de zorg werkelijk is verslechterd.

III Toezicht

De kwaliteitsindicatoren kunnen deel uitmaken van een grotere set van informatiebronnen waarmee interne en externe toezichthouders toezicht houden op de kwaliteit van zorg. Daarbij kunnen kwaliteitsindicatoren helpen om een oordeel te vormen over de kwaliteit van zorg, of als screeningsinstrument om een selectie te maken van zorgaanbieders die nader onderzocht zullen worden.

Inhoudsvaliditeit

Toezichthouders hanteren mogelijk een definitie van kwaliteit van zorg die afwijkt van de sectorspecifieke kwaliteitsvisie en hebben veelal speciale aandacht voor veiligheid. Daarnaast hebben toezichthouders veelal kennis van het primaire proces waardoor zij naast uitkomsten ook aandacht hebben voor processen en structuren. Toezichthouders maken daarom gebruik van zowel uitkomst-, als proces- en structuurindicatoren. Van proces- en structuurindicatoren is duidelijk dat

ze beïnvloedbaar zijn door zorgaanbieder(s). Daarnaast leveren deze indicatoren informatie over de wijze waarop de zorg kan worden verbeterd.

Vanwege de verschillen tussen de perspectieven is het oordeel over de inhoudsvaliditeit op setniveau voor een toezichthouder niet direct bruikbaar. Een set uitkomstindicatoren die een volledig beeld geeft van de kwaliteit van zorg kan voor toezichthouders onvoldoende informatie bieden om hun taak uit te voeren. De beschikbare informatie per indicator is naar verwachting wel voldoende voor toezichthouders om zelf te oordelen over de inhoudsvaliditeit op setniveau. Oordelen op indicatorniveau komen beter overeen. In de regel zullen indicatoren die vanuit het patiëntkeuzeperspectief inhoudsvalide zijn, ook voor toezichthouders bruikbare informatie opleveren.

Vertekening: populatievergelijkbaarheid

Populatieverschillen leiden bij gelijke kwaliteit tot verschillen in indicatorwaarden. Correctie voor deze verschillen is zowel voor patiënten als voor toezichthouders wenselijk. Daarbij zullen toezichthouders, vanuit hun informatiebehoefte, in principe voor dezelfde set van factoren corrigeren. Voor toezichthouders heeft correctie echter ook een nadeel indien de geaggregeerde resultaten alleen gecorrigeerd worden gepresenteerd. In dat geval is het lastiger om de zorg die geleid heeft tot de indicatorwaarden te traceren. Het kan dan bijvoorbeeld lastig zijn om vast te stellen wat het exacte aantal incidenten was tijdens een meetperiode.

Indien de correctie wordt uitgevoerd met empirical Bayes wordt de bepaling van het aantal incidenten nog een stuk lastiger. In dat geval krimpt (*schrinkage*) bij slecht presterende instellingen elke meting naar het gemiddelde. Daardoor kunnen onveilige situaties vele jaren onopgemerkt blijven. Hoewel het mogelijk is de empirical Bayes krimp bij benadering ongedaan te maken, zal het lastig blijven het exacte aantal incidenten te bepalen. Daarom hebben toezichthouders belang bij zowel gecorrigeerde als ongecorrigeerde geaggregeerde gegevens.

Correctie voor populatieverschillen is voor toezichthouders minder noodzakelijk dan voor patiënten. Toezichthouders zijn beter in staat de populatieverschillen zelf te verdisconteren en daarmee indicatorwaarden vergelijkbaar te maken. Toezichthouders hoeven daarom weinig waarde toe te kennen aan het deskundigenoordeel over de populatievergelijkbaarheid. Indicatoren die als niet voldoende vergelijkbaar worden beoordeeld zijn voor toezichthouders mogelijk wel bruikbaar.

Vertekening: registratievergelijkbaarheid

Indien de registratie van alle gegevens die nodig zijn om een indicatorwaarde te bepalen niet vergelijkbaar is, dan is de indicator slecht bruikbaar als patiëntkeuze-informatie. Ook voor toezichthouders zullen gegevens minder bruikbaar zijn. Het deskundigenoordeel over de registratievergelijkbaarheid is daardoor ook voor toezichthouders relevant. Wel hebben toezichthouders vaak mogelijkheden om indicatorwaarden te vergelijken met andere bronnen van informatie, waardoor zij in staat zijn de kwaliteit van de registratie te beoordelen.

Vertekening: steekproefvergelijkbaarheid

Indien niet alle patiënten betrokken zijn bij de meting, dan is het mogelijk dat de indicatorwaarden niet representatief zijn voor de geleverde kwaliteit van zorg. Gebrekkige representativiteit van de resultaten zal ook toezichthouders hinderen bij hun werkzaamheden, en het deskundigenoordeel is daarmee ook voor toezichthouders bruikbaar. Toezichthouders hebben echter mogelijkheden om andere bronnen van informatie te gebruiken en zo na te gaan of de resultaten representatief zijn.

Statistisch betrouwbaar onderscheiden

Toeval is ook voor toezichthouders relevant. Het criterium statistisch betrouwbaar onderscheiden wordt voor het patiëntkeuzeperspectief echter beoordeeld per indicator en per jaar. Indien toezichthouders hun oordelen slechts baseren op afzonderlijke indicatoren gemeten op een enkel

meetmoment, dan zijn de oordelen over statistisch betrouwbaar onderscheiden ook voor de toezichthouders direct bruikbaar.

Het is echter waarschijnlijk dat toezichthouders indicatoren die zorginhoudelijk samenhangen gezamenlijk gebruiken om tot een oordeel te komen. Het kan daarbij gaan om meerdere meetmomenten van dezelfde indicator, of meerdere indicatoren die zorginhoudelijk samenhangen of combinaties van indicatoren en andere bronnen van informatie. In dat geval kunnen afzonderlijke indicatoren die allen individueel statistisch niet voldoende kunnen worden onderscheiden, bijdragen aan een gezamenlijk oordeel: een samengestelde indicator. Deze kan statistisch wel voldoende worden onderscheiden.

Indien afzonderlijke indicatoren worden gecorrigeerd met empirical Bayes, dan wordt het ontwikkelen van samengestelde indicatoren lastig. De krimp moet voor dat doel ongedaan worden gemaakt. Het ontwikkelen van betrouwbaarheidsintervallen is in dat geval ook lastig. Indien toezichthouders gebruik wensen te maken van samengestelde indicatoren dan zouden deze idealiter worden gebaseerd op ruwe data op patiëntniveau (met maatregelen om de privacy van patiënten te borgen).

Concluderend

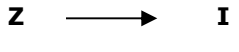
Indicatoren die zijn ontwikkeld vanuit een patiëntkeuzeperspectief zijn in veel gevallen minder bruikbaar voor andere gebruikersdoelen. Voor deze doelen zijn aanvullende bewerkingen nodig. De correctie voor populatieverschillen in het algemeen en correcties op basis van empirical Bayes in het bijzonder, hinderen het gebruik van de data voor andere gebruiksdoelen. De verzamelde data kunnen echter pas optimaal worden gebruikt voor alle gebruiksdoelen indien aanvullende analyses van de ruwe data op patiëntniveau mogelijk zijn. Die aanvullende analyses vereisen niet dat alle gebruikers zelf toegang krijgen tot de ruwe data.

Bijlage 4: Schematische toelichting vertekening

Directed Acyclical Graphs

De indicatorstandaard is gebaseerd op een indeling van vertekening en toeval die volgt uit de epidemiologie. De begrippen komen soms niet overeen met begrippen die in andere wetenschapsgebieden worden gehanteerd, en ook binnen de epidemiologie worden verschillende definities naast elkaar gehanteerd. Dit leidt vaak te verwarring. Om de verwarring tot een minimum te beperken worden de concepten hieronder toegelicht aan de hand van figuren.

Deze figuren worden Directed Acyclical Graphs, of kortweg DAG's genoemd. Uit de naam blijkt dat een effect maar een richting heeft (directed). Daarnaast kan een effect niet cyclisch zijn, dat wil zeggen een effect kan zichzelf niet veroorzaken en cirkels zijn dus onmogelijk. In de DAG's die hieronder worden gebruikt is een denkbeeldige tijdslijn opgenomen die van links naar rechts loopt. Figuur B6.1 geeft een voorbeeld van een DAG. In dit voorbeeld loopt de pijl van Z (zorgaanbieder) naar I (indicator). Dit betekent dat Z verschillen in I heeft veroorzaakt. Indien dit schema de werkelijkheid beschrijft, dan veroorzaakt Z verschillen in I.

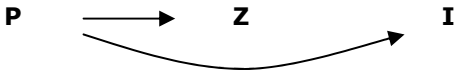


Z *Zorgaanbieder*
I *Indicator*

figuur B6.1 Het zorgaanbieder veroorzaakt variatie in een indicator

Populatievergelijkbaarheid

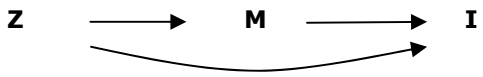
Figuur B6.2 geeft het vertekenende effect van populatiekenmerken (P) op de geschatte relatie tussen zorgaanbieder (Z) en kwaliteit van zorg (K) schematisch weer. Indien bijvoorbeeld mobiele patiënten minder kans maken op decubitus (P beïnvloedt K) en de mobiliteit van patiënten die opgenomen worden bij zorgaanbieder A beter is dan die van patiënten die opgenomen worden bij zorgaanbieder B (P beïnvloedt Z) en de mobiliteit van patiënten bij opname niet beïnvloedbaar is (P gaat vooraf aan opname Z) dan zal A bij gelijke zorgkwaliteit beter scoren dan B.



Z *Zorgaanbieder*
I *Indicator*
P *Populatiefactor*

figuur B6.2 Het vertekenende effect van populatiekenmerken op de geschatte relatie tussen zorgaanbieder en een indicator

In figuur B6.3 worden deze met M aangeduid. Correctie voor mediërende factoren is onwenselijk. Bijvoorbeeld, zorgaanbieder A past bij immobiele patiënten vaker wisselligging (M) toe dan zorgaanbieder B, en wisselligging leidt tot minder decubitus. Na correctie voor wisselligging zouden A en B gelijk scoren. Correctie voor wisselligging (M) is ongewenst, omdat het juist de geleverde zorg van de organisatie is die ervoor zorgt dat er decubitus minder voorkomt.

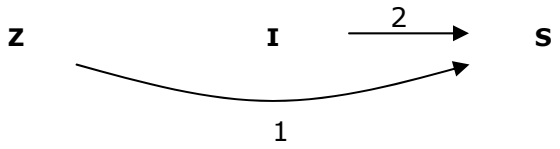


Z *Zorgaanbieder*
I *Indicator*
M *Mediërende factor*

figuur B6.3 Invloed van mediërende factor op een indicator

Steekproefvergelijkbaarheid

Steekproefvertekening ontstaat wanneer het selectief includeren van personen in een onderzoek leidt tot vertekening van de geschatte kwaliteit. Deze vertekening treedt op als zowel de zorg-aanbieder als de uitkomst van zorg de selectie van waarnemingen (de steekproef) beïnvloeden. Figuur B6.4 illustreert hoe dit werkt voor indicatorvergelijkingen. Z staat voor de zorgaanbieder, I voor de waarden van de indicator(set), en S voor de factoren of criteria waarop geselecteerd is.



Z Zorgaanbieder
I Indicator
S Factoren waarop geselecteerd is

figuur B6.4 Invloed van steekproefvertekening

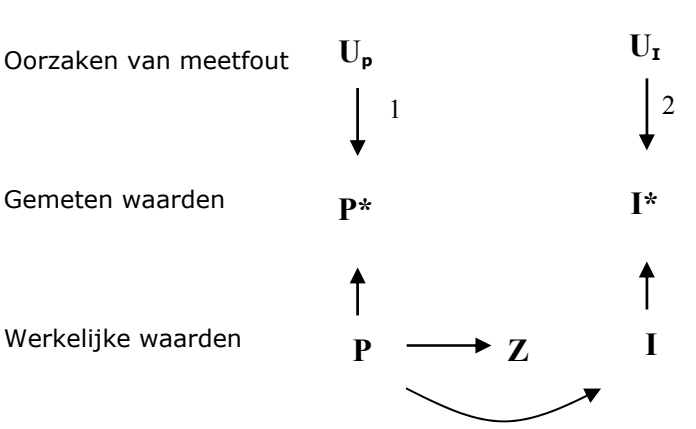
Registratievergelijkbaarheid: Typologie meetfouten

Fouten in gegevens kunnen daarin terecht komen als iets misgegaan is met het meten, registreren, opslaan, en/of verwerken van de gegevens. Al deze fouten noemen wij meetfouten. Vrijwel geen enkele meting of registratie is vrij van meetfouten. Bij indicatorvergelijkingen kunnen fouten optreden in de meting van indicatoren en/of populatievariabelen. Er zijn drie soorten meetfouten:

1. onafhankelijke meetfout
2. afhankelijke meetfout
3. differentiële meetfout

Onafhankelijke meetfout

Onafhankelijke meetfout in een variabele is meetfout die op geen enkele manier samenhangt met andere variabelen of meetfouten in andere variabelen. Oftewel, onafhankelijke meetfouten hangen onderling niet samen. Figuur B6.5 is een schematische weergave van onafhankelijke meetfouten in een indicatorvergelijking. Het achterliggende idee is dat zorgaanbieders een bepaalde kwaliteit van zorg leveren en dat die indicatorwaarden vergeleken kunnen worden na correctie voor verschillen in behandelpopulatie. (Zie hoofdstuk 3 voor een bespreking van de noodzaak voor correctie van verschillen in behandelpopulatie.) Die verschillen in behandelpopulatie noemen we populatie-factoren.



Z de zorginstelling,
I werkelijke waarden van een indicator(set),
I de gemeten waarden van een indicator(set),*
P de werkelijke waarden van de verstorende populatiekenmerken,
P de gemeten waarden daarvan,*
UP onbekende (unknown) factor die alleen meetfout geeft in de bepaling van de verstorende factoren, te weten verschillen in behandelpopulatie,
UI onbekende (unknown) factor die alleen meetfout geeft in de bepaling van de indicatoren(sets).

figuur B6.5 Onafhankelijke meetfout

Indien een dataset onafhankelijke meetfouten bevat voor zowel de populatiekenmerken als de indicatorwaarden dan kan dat worden weergegeven in figuur B6.4. In deze figuur komt dit tot

uiting doordat de gemeten verstorende populatievariabelen P* niet alleen bepaald worden door de werkelijke populatiekenmerken P, maar ook door een (vaak onbekende) factor U_V. De gemeten indicatorwaarden I* worden bepaald door de werkelijke indicatorwaarde I en een factor U_I. Deze meetfouten zijn onafhankelijke meetfouten omdat zij los staan van de (meting) van de andere factoren. Bijvoorbeeld, de meetfout in P* is onafhankelijk van Z, I, I* en U_I. Samengevat: onafhankelijk meetfout treedt op als aan *ten minste* één van de volgende twee voorwaarden is voldaan:

- een onbekende factor de meting van verstorende populatievariabelen beïnvloedt (pijl 1);
- een andere onbekende factor de meting van de indicatorwaarden beïnvloedt (pijl 2).

Onafhankelijke meetfouten kunnen het gevolg zijn van toeval of slordigheid. In de wetenschapspraktijk wordt wel eens getest in hoeverre de meting (on)afhankelijk is van de codeur. De zogenaamde inter-rater betrouwbaarheid van een meetinstrument is bij voorbeeld hoger naarmate een herhaalde meting door dezelfde codeur in dezelfde periode en setting vaker dezelfde waarde oplevert als de eerste meting.

Deze onafhankelijke meetfouten kunnen de indicatorvergelijkingen vertekenen indien de meetfout betrekking heeft op de populatie variabelen. Dergelijke meetfouten leiden tot een onderschatting van daadwerkelijke effecten van populatie op de indicatorwaarde. De geschatte verschillen gaan in de richting van 'de nul' (geen verschil), vandaar dat deze bron van vertekening attenuatie (verdunnings- of vervlakkings-) vertekening wordt genoemd[19]. Onderschatting van de populatie-effecten leidt tot ondercorrectie voor populatieverschillen en werkt dus uit in het voordeel van zorgaanbieders met een gunstige populatie. Zij krijgen bij gelijke kwaliteit betere uitkomsten. Indien de werkelijke populatiewaarde niet goed te observeren is, dan is ook de prikkel tot het selecteren van gunstige risico's door de zorgaanbieder beperkt.

Onafhankelijke meetfouten in de indicatorwaarden leiden *niet* tot vertekening maar tot een lagere statistische betrouwbaarheid en een verlies van onderscheidingsvermogen (zie hoofdstuk 6).

Afhankelijke meetfouten

Het tweede type meetfout is de afhankelijke meetfout. Eén bepaalde variabele anders dan de populatiekenmerken, zorgaanbieders, en indicatoren veroorzaakt fouten in de meting van de populatiekenmerken *en* indicatorwaarden. In andere woorden, de meting van zowel indicator als populatievariabelen is *afhankelijk* van eenzelfde derde factor. Figuur B6.6 representeert dergelijke afhankelijke meetfout. De datasets met populatiekenmerken P* en indicatoren waarden I* bevatten meetfout die veroorzaakt wordt door U_{PI}. Afhankelijke meetfout treedt dus op als aan de volgende voorwaarden beide wordt voldaan:

- een factor beïnvloedt de meting van verstorende variabelen (pijl 1);
- diezelfde factor beïnvloedt de meting van en indicatoren (pijl 2).

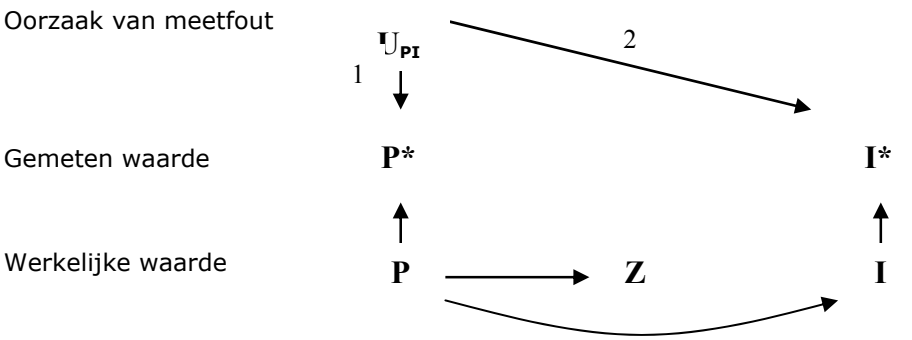
Afhankelijke meetfouten in de indicatoren en populatievariabelen leiden tot vertekening van de indicatorvergelijking. Het effect van deze vertekening op de indicatorvergelijking bestaat deels uit de hierboven beschreven ondercorrectie voor populatieverschillen, en deels uit een samengesteld effect. Dat samengestelde deel kan zorgen voor extra vervlakking, maar kan de vervlakking ook laten afnemen[20]. Hierdoor is de richting van het totale effect lastig te voorspellen.

Afhankelijke meetfout kan bijvoorbeeld ontstaan wanneer de metingen worden verricht door medewerkers met uiteenlopende diagnostische vaardigheden. Zo mist een beter getraind oog minder aandoeningen en dat heeft consequenties voor zowel de indicatorwaarden als de populatiekenmerken.

---

[19] In de wetenschappelijke literatuur staat deze vorm van bias bekend als attentuation, regression dilution, (classical)errors-in-variables

[20] Afhankelijk van het teken van de samenhang in pijl 1 en pijl 2.



Z de zorginstelling, I werkelijke waarden van een indicator(set), I de gemeten waarden van een indicator(set), P de werkelijke waarden van de verstorende populatiekenmerken, P* de gemeten waarden daarvan, U_{PI} meetfout die veroorzaakt wordt door U_{PI}*

figuur B6.6 Afhankelijke meetfout

Differentiële meetfout

Tot slot, een meetfout kan ook optreden als de onderzochte factoren P, Z en I zelf invloed hebben op de metingen. Een zorgaanbieder kan de meting van de populatievariabelen beïnvloeden. Populatiemeetfouten verschillen, of *differentiëren*, dan per zorgaanbieder. Daarnaast kan ook de werkelijke indicatorwaarde de meting van de populatievariabelen bepalen. Figuur B6.5 representeert deze situatie. De zorgaanbieder en de werkelijke indicatorwaarde beïnvloeden de gemeten waarden van de populatiekenmerken P*. Differentiële meetfout in populatievariabelen treedt op als aan tenminste één van de volgende voorwaarden is voldaan:

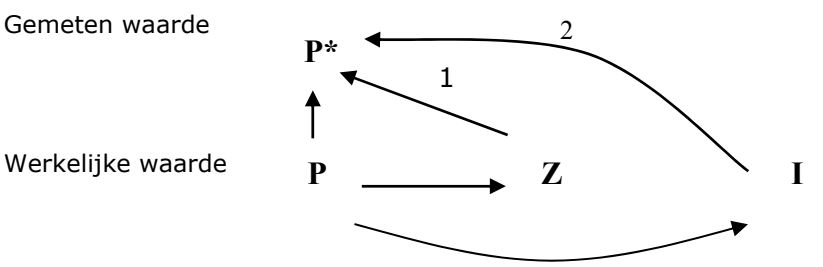
- de zorgaanbieder de meting daarvan beïnvloedt (pijl 1);
- de werkelijke indicator waarde dat doet (pijl 2).

Een voorbeeld van pijl 1 is een zorgaanbieder die de populatievariabelen slecht registreert, of vanuit strategisch gedrag de ongustige populatiekenmerken aandikt. Een voorbeeld van pijl 2 is herinneringsvertekening (recall bias). Deze vorm van vertekening ontstaat wanneer bij de herinnering zelf afhankelijk is van de uitkomst van de indicator. Stel bijvoorbeeld dat iemand decubitus heeft. In dat geval is de kans waarschijnlijk groter dat behandelaars zich populatiekenmerken herinnerden die bijdragen aan het ontstaan van decubitus, dan bij patiënten die geen decubitus ontwikkelden.

Op dezelfde manier kan differentiële meetfout optreden in de meting van de indicatorwaarden. De werkelijke waarden van de populatiekenmerken en/of de zorgaanbieders beïnvloeden de meting van de indicatorwaarden. Hiervan was het voorbeeld in de inleiding een illustratie, waarin het ene verpleeghuis bij twijfel over de graad van decubitus een hogere graad toekende, en het andere verpleeghuis juist een lagere graad. Differentiële meetfout in het vaststellen van de indicatorwaarden wordt weergegeven in figuur B6.7. De figuur geeft aan dat differentiële meetfout in de indicatorwaarde optreedt als aan tenminste één van de volgende voorwaarden is voldaan:

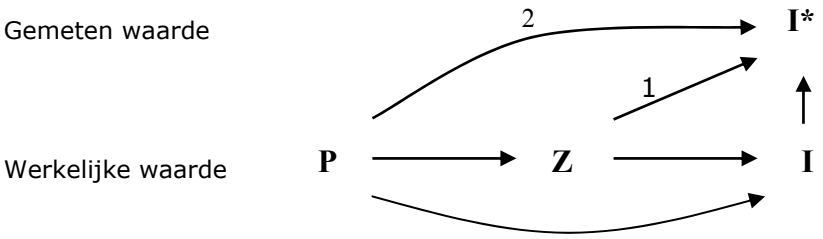
- de zorgaanbieder de meting daarvan beïnvloedt (pijl 1);
- een (set van) populatiekenmerken dat doet (pijl 2).

Pijl 1 kan zich voordoen bij zorgaanbieders die de indicatorwaarden strategisch invullen. Het is daarom de meest gevreesde vorm van meetfout. Deze vorm van meetfout heeft een sterk en direct effect op de indicatorvergelijking. Het is voor zorgaanbieders duidelijk dat zij beter uit de indicatorvergelijking komen wanneer zij de indicatorwaarden in hun voordeel aanpassen. Een voorbeeld van pijl 2 is onderdiagnostiek van depressie bij ouderen. Bij oudere mensen worden somberheid en interesseverlies vaak beschouwd als een natuurlijke, niet pathologische, reactie op ziekte, terwijl er volgens de psychiatrische criteria wel een depressie kan zijn.



P de werkelijke waarden van de verstorende populatiekenmerken,
P de gemeten waarden daarvan,*
Z de zorginstelling,
I werkelijke waarden van een indicator(set)

figuur B6.7 Differentiële meetfout in verstorende variabelen



P de werkelijke waarden van de verstorende populatiekenmerken,
Z de zorginstelling,
I werkelijke waarden van een indicator(set),
I de gemeten waarden daarvan*

figuur B6.8 Differentiële meetfout in verstorende variabelen

Typologie en correctie voor meetfouten

Voor vertekening door onafhankelijke meetfouten kan gecorrigeerd worden als de vertekening zelf kan worden gemeten. Dit kan bijvoorbeeld door de gemeten verstorende variabelen te vergelijken met een gouden standaard. De aanname is dat de gouden standaard een perfecte weergave is van de werkelijke populatiekenmerken (case-mix). Omdat deze aanpak arbeidsintensief is, wordt in de praktijk veelal gebruik gemaakt van een set van veronderstellingen waarmee de vervlakking kan worden gecorrigeerd. Indien die veronderstellingen opgaan kan gebruik worden gemaakt van de zogenaamde *regression dilution ratio.* Zonder aanvullend onderzoek is het lastig te bepalen of de aannamen voldoende realistisch zijn.

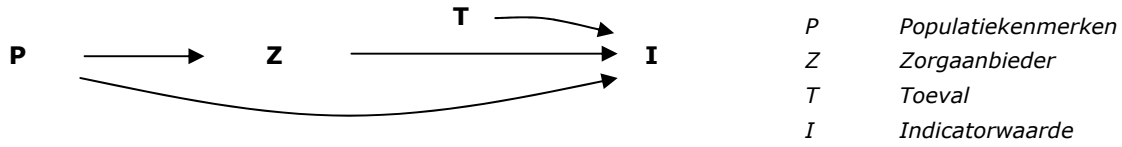
Om afhankelijke meetfout te corrigeren is ook kennis nodig over de factoren die meetfout veroorzaken. Dit wordt duidelijk in figuur B6.6: corrigeren voor U_{PI} heft de associatie tussen P* en I* op en daarmee het vertekende effect van deze associatie op vergeleken prestaties Z*-I*. Correctie voor dit soort bias vereist nader onderzoek zodat bekend wordt wat die factoren U_{PI} zijn. Hiervoor kunnen onder andere ankervignetten gebruikt worden. Ankervignetten beschrijven zorg van een bepaalde kwaliteit. Wanneer deze strak omschreven zorg verschillend wordt beoordeeld door verschillende groepen, dan kunnen de verschillen gebruikt worden voor een correctie. Als het aantal factoren, of groepen factoren, die meetfout veroorzaken heel groot is, dan kan het nodig zijn om het ankervignet te gebruiken bij elke meting. Hierdoor zullen de kosten van een meting sterk stijgen.

In het geval van meting van geleverde zorg is het waarschijnlijk dat de meting per zorgaanbieder afwijkt als de zorgaanbieder de meting zelf uitvoert (Lu en Ma 2002). Dit leidt tot ten minste zoveel groepen zijn als er zorgaanbieders zijn. Hierdoor is het bepalen van een gouden standaard in de praktijk onhaalbaar. Daarnaast kan eventueel strategisch gedrag van de zorgaanbieder deze vorm van correctie teniet doen.

Samenvattend kan worden gesteld dat er statistische technieken zijn om te corrigeren voor elke vorm van meetfout. Toch is het gebruik van deze technieken te vergelijken met genezen waar voorkomen beter was geweest. Indien voorkomen onmogelijk of duur is, dan kunnen statistische technieken in sommige gevallen een nuttige bijdrage leveren. In alle gevallen zijn statistische technieken nuttig vormen van informatiebias te detecteren, en om een inschatting te geven van de ernst van de vertekening.

Toeval

In figuur B6.9 staat P voor populatiekenmerken, Z voor zorgaanbieder, T voor toeval en I voor indicatorwaarde. Uit de figuur blijkt dat de populatiekenmerken (P) structureel samenhangen met de zorgaanbieder en daarnaast ook direct met de waarde van een indicator. Toeval bepaalt ook de waarde van de indicator, maar de invloed van het toeval hangt niet structureel samen met de zorgaanbieder. De invloed van toeval op de indicatorwaarde is te verkleinen door het aantal waarnemingen te verhogen: de toevallige afwijkingen heffen elkaar dan min of meer op. In de praktijk lukt het vaak niet om grote hoeveelheden waarnemingen te realiseren, waardoor de indicatorwaarde van een zorgaanbieder wel wordt beïnvloed door toeval. De toevallige gezondheidstoestand van de patiëntenpopulatie tijdens de meetmomenten kan dan dus bepalend zijn voor de indicatorwaarden.



figuur B6.9 Schematische weergave van toeval

Bijlage 5: Triangulatie, construct en criterium validiteit

Triangulatie

Triangulatie wordt in de sociale wetenschappen gebruikt om aan te geven dat meerdere technieken worden gebruikt zodat de resultaten te voorspellen zodat de resultaten kunnen worden vergeleken en gecontroleerd. Dit wordt ook "cross examination" genoemd. Bij het registreren kan triangulatie gebruikt worden door de informatie aan zowel de zorgprofessional als de patiënten uit te vragen. Zo wordt een indruk verkregen van de overeenstemming in de (indruk van) de aangeboden zorg en de ervaren zorg.

Constructvaliditeit

Constructvaliditeit gaat over de vraag of de resultaten van een onderzoek een indicatie zijn voor het construct waarover je een uitspraak wilt doen. Resultaten van een onderzoek kunnen aansluiten bij de theorie, maar er kan onvoldoende rekening zijn gehouden met andere factoren die ook invloed hebben op het onderzochte construct. Constructen zijn daarbij theoretische begrippen met ideale eigenschappen.

De 'construct validiteit' van een indicator is hoog als er een sterke positieve associatie is tussen de waarde van de betreffende indicator en de waarde van een andere indicator die het onderzochte kwaliteitsaspect betreft of een wetenschappelijke studie die deze associatie bevestigd. "[The indicator] should be related to other indicators intended to measure the same or related aspects of quality"(AHRQ, 2010).

Constructvaliditeit is verder te onderscheiden in convergente- en divergente validiteit. Convergente validiteit beschouwt de samenhang tussen de resultaten van het oorspronkelijke onderzoek en de resultaten van een gelijksoortig onderzoek. Hoe hoger de correlatie, hoe meer valide de test. Er kan ook gekeken worden naar de samenhang tussen de resultaten van onderzoek en observeerbaar gedrag. Bij divergente validiteit wordt gekeken naar de samenhang tussen de resultaten van het oorspronkelijke onderzoek en de resultaten van een ander onderzoek. Echter, hier geldt dat de correlatie zoveel mogelijk rond het nulpunt moet liggen, voor een meer valide test.

Criteriumvaliditeit

Criteriumvaliditeit beschrijft in welke mate een test voorspellende waarde heeft. Criteriumvaliditeit is verder te onderscheiden in predictieve validiteit en concurrent validity. Predictieve validiteit heeft betrekking op de vraag in hoeverre een test kan voorspellen wat het in theorie moet kunnen voorspellen. Kun je naar aanleiding van een testscore voorspellen hoe de participant zich in de werkelijkheid gaat gedragen? *Concurrent validity* beschouwt in hoeverre de resultaten correleren met gelijktijdig beschikbare criteriumgegevens. De 'criterium validiteit' van een indicator is hoog als er een sterk positieve associatie bestaat tussen de uitkomst van de indicator en de uitkomst van een meting betreffende hetzelfde kwaliteitsaspect waarvan de validiteit reeds is vastgesteld.

Onderscheid construct- en criteriumvaliditeit

Het onderscheid tussen beide soorten validiteit is subtiel: in het eerste geval (construct validiteit) gaat het om de vergelijking van twee maten die beide tot doel hebben de onderliggende kwaliteit te meten. Het vinden van een hoge associatie tussen de twee maten is dan een indicatie dat de maten inderdaad deze onderliggende kwaliteit lijken te meten. Het is hierbij zinvol om te onderzoeken in hoeverre de scores (de indicatorwaarden) van de indicatoren uit een zelfde set met elkaar samenhangen. Gedurende de doorontwikkeling van een indicatorset dient ten minste aandacht te worden besteed aan de afwezigheid van dergelijke associaties, daar waar de betreffende indicatoren wel aan hetzelfde kwaliteitsaspect zouden dienen te refereren. De

afwezigheid van deze associaties is dan een aanwijzing voor een mogelijk gebrek aan validiteit van de afzonderlijke indicatoren.

In het tweede geval (criterium validiteit) gaat het om de vergelijking tussen een (nieuwe) indicator en een 'gouden standaard' (het 'criterium') waarvan de validiteit vaststaat, of als a priori wordt verondersteld.

Bijlage 6: Directe versus indirecte standaardisatie

Corrigeren voor verschillen in de behandelpopulatie kan met vele verschillende methoden. De twee belangrijkste groepen van methoden zijn directe en indirecte standaardisatiemethoden. Bij directe standaardisatie worden de uitkomsten in de behandelpopulatie vergelijkbaar gemaakt aan de uitkomsten in de referentiepopulatie met behulp van gewichten. De referentiepopulatie is bijvoorbeeld gelijk aan alle behandelpopulaties van alle zorgaanbieders, maar kan ook gelijk zijn aan de totale verzekerde populatie waar een zorgverzekeraar bepaalde zorg voor inkoopt.

Bij directe standaardisatie worden populatieverschillen ongedaan gemaakt door waarnemingen te wegen in de analyse. Indien mannen 75% van de behandelpopulatie uitmaken en slechts 50% van de referentiepopulatie dan krijgen de mannen in de behandelpopulatie een gewicht van 2/3 en vrouwen een gewicht van 2. Bij het bepalen van het gewogen gemiddelde van de behandelpopulatie tellen mannen 2/3x mee en vrouwen tellen 2x mee.

Het bepalen van de gewichten wordt lastiger wanneer het aantal populatiekenmerken waarvoor gecorrigeerd wordt toeneemt. In dat geval kan gebruik gemaakt worden van parametrische technieken zoals mogelijk is bij inverse probability weighting (IPW). Deze gewichten zijn gelijk aan de inverse van de kans om geobeserveerd te worden. De kan om geobserveerd te worden kan worden geschat met een parametrisch (regressie)model (Hernan en Robins, 2006).

Bij indirecte standaardisatie worden de uitkomsten van de behandelpopulatie vergeleken met de verwachte uitkomsten op basis van de uitkomsten die patiënten met vergelijkbare karakteristieken hadden in de referentiepopulatie. Daarbij wordt een verwacht gemiddelde uitgerekend dat dient om de uitkomsten van de behandelpopulatie te vergelijken met de uitkomsten in de referentiepopulatie.

Regressie

De technieken die hierboven beschreven staan zijn niet parametrisch: zij maken geen gebruik van de voor regressiemodelen benodigde veronderstellingen. Daardoor zijn deze technieken inefficiënt en hebben beide vele waarnemingen per zorgaanbieder nodig. Daarom zijn voor zowel directe als indirecte standaardisatiemethoden parametrische varianten ontwikkeld.

De ontwikkeling van parametrische varianten van directe standaardisatie, zoals vormen van Inverse Probability Weighting (IPW), is echter recent van aard. Er is nog weinig ervaring met deze groep van technieken voor het vergelijkbaar maken van de kwaliteit van zorgaanbieders. Daarom is het op dit moment gebruikelijk om indirecte standaardisatie met behulp van regressie te gebruiken..

Voor- en nadelen

Het is mogelijk dat directe en indirecte standaardisatie tot tegenstrijdige uitkomsten leiden. Daarbij beantwoordt directe standaardisatie de vraag welke uitkomst de zorgaanbieder zou hebben gehad indien de populatie vergelijkbaar was met het landelijk gemiddelde. Indirecte standaardisatie beantwoordt de vraag hoe de zorgaanbieder presteerde ten opzichte van de verwachte uitkomst.

Directe standaardisatie sluit daarbij aan bij de patiënt of zorginkoper die op het punt staat te kiezen tussen zorgaanbieders. Directe standaardisatie kent echter ook een nadeel. Bij indirecte standaardisatie telt elke patiënt even zwaar mee, terwijl bij directe standaardisatie de uitkomsten van patiënten die relatief weinig behandeld worden veel zwaarder wegen. Directe standaardisatie

bevat daardoor een prikkel voor een zorgaanbieder om zich te richten op de patiënten die relatief weinig worden behandeld.

In de praktijk wordt het meerendeel van de populatievergelijkbaarheidscorrecties gebaseerd op indirecte standaardisatie en uitgevoerd met behulp van (emprical Bayes) regressie. Deze technieken vereisen minder waarnemingen en zijn gemakkelijker uit te voeren dan de andere beschikbare technieken. Directe standaardisatie zal in de nabije toekomst waarschijnlijk een haalbaar alternatief worden, maar er is op dit moment internationaal nog weinig ervaring opgedaan met deze technieken voor de vergelijking van kwaliteit tussen zorgaanbieders.

Bijlage 7: Relatieve en absolute uitkomstmaten

Er bestaan vele verschillende uitkomstmaten voor indicatoren. Het belangrijkste onderscheid tussen groepen van uitkomstmaten is het verschil tussen relatieve en absolute maten. Relatieve maten drukken de kwaliteit uit in een verhoudingsgetal. Een voobeeld is de verhouding tussen ziekenhuissterfte in een ziekenhuis vergeleken met de ziekenhuissterfte in alle Nederlandse ziekenhuizen. Indien de uitkomst 1,1 is dan heeft het ziekenhuis 10% meer sterfgevallen dan het landelijke gemiddelde. Indien gewenst kan deze relatieve uitkomstmaat gecorrigeerd worden voor verschillen in behandelpopulatie.

Absolute maten drukken de verschillen uit in het verschil in percentage. Zo kan de situatie hierboven overeenkomen met een landelijk gemiddelde ziekenhuissterfte van 2% en kan het genoemde ziekenhuis een gemiddelde sterftekans hebben van 2,2%. Ook deze maat kan worden gecorrigeerd voor verschillen in behandelpopulatie.

Een keuze tussen absoluut en relatief is vaak subjectief. Beide uitkomstmaten vertellen een ander verhaal en beide zijn in veel gevallen interessant. In medisch onderzoek is de relatieve maat het meest populair. Zo is het redelijk om te veronderstellen dat de extra kans op sterfte van een hoog risicopatiënt niet 0,2%-punt is, maar 10% meer dan de kans bij een gemiddeld ziekenhuis. Indien de hoog risico patiënt gemiddeld 50% kans op overlijden heeft, dan is dat in het bovenstaande ziekenhuis mogelijk 55%. En voor laag risicopatiënten met een kans van 0,01% zal waarschijnlijk gelden dat hun risico verhoogt tot 0.011%. Een minieme toename van 1 op de 100.000 patiënten.

Een nadeel van relatieve schalen is dat de relatieve verhouding vaak niet mooi constant blijft en dat het veel uitmaakt of sterfte of juist overleving wordt gekozen als uitkomst. Ook zijn er vele mogelijke relatieve uitkomstmaten waaruit gekozen kan worden. Omdat het patiëntkeuzeperspectief leidend is, ligt het voor de hand een relatief risico te kiezen vanwege de eenvoudige interpretatie. Dat relatief risico kan vervolgens zo worden samengesteld dat de relatieve verhouding zoveel mogelijk constant blijft.

Een absolute maat heeft ook voordelen. Zo kan een absolute maat zo gekozen worden dat direct blijkt hoe hoog het percentage sterfgevallen is (2,2%). Die hoogte is relevante als patiëntkeuze-informatie want een relatieve verhoging van 10% op een kans van 0.01% is niet wezenlijk.

De berekening van de signaalvlag statistisch betrouwbaar onderscheiden is gebaseerd op een relatieve uitkomstmaat. Voor absolute uitkomstmaten geeft de signaalvlag een indicatie van de invloed van toeval. Ook het ontwikkelen van samengestelde indicatoren is afhankelijk van de keuze van uitkomstmaat. Absolute maten zijn eenvoudiger te verwerken in een samengestelde indicator.

Bijlage 8: Weergave indicatoren

1. [titel indicator]	
Indicator	
Relatie tot kwaliteit	[Wat is de relatie van de indicator tot de kwaliteit van zorg en/of het visiedocument]
Operationalisatie	[beschrijving operationalisatie]
	[beschrijving teller]
Noemer	[beschrijving noemer]
Definitie(s)	[uitleg relevante definities]
In/ exclusiecriteria	[beschrijving in- en exclusiecriteria]
Beschrijving popualtieverschillen	[Indien hiervoor gecorrigeerd dient te worden: beschrijving potentiële factoren op patiëntniveau, die score kunnen beïnvloeden. Inclusief onderbouwing = bron]
Bronnen tbv inhoudsvaliditeit	[beschrijving (Nederlandse) bron, met indien mogelijk level of evidence]
Gegevensverzameling	
Bron	[beschrijving bron zorgaanbieder, zo nodig voor teller en noemer apart]
Verslagjaar	[periode waarover gegevens worden aangeleverd]
Rapportagefrequentie	…x per verslagjaar
Meetniveau	[praktlijklokatie, specialisme, zorgverlener]
Ordeningskader	
Type indicator	**Structuurindicatoren** betreffen vragen naar de organisatorische randvoorwaar-den van de zorg (o.a. aanwezigheid van registraties, kwalificaties van personeel) **Procesindicatoren** betreffen metingen aan de procesgang van de zorg (o.a. wacht- en doorlooptijden, volume indicatoren, indicatiestelling, richtlijnconformiteit) **Uitkomstindicatoren** betreffen metingen van de uitkomsten van de geleverde zorg (o.a. patiënttevredenheid, 10-jaars overleving, mate van ziekteactiviteit, kwaliteit van leven, complicaties, sterfte). Een onderverdeling kan hier nog worden gemaakt naar 'intermediate outcomes' en echte 'eindpunten' van zorg. In het eerste geval gaat het bijvoorbeeld om een 'gecontroleerde bloeddruk' (de medische parameters aan het einde van de behandeling[21] zijn goed); in het laatste geval om bijvoorbeeld het 'aantal jaren zonder cardiovasculair incident'.
Kwaliteitsdomein	De kwaliteitsdomeinen volgen uit de sectorspecifieke kwaliteitsvisie en dienen per domein te worden beschreven.
Fase in zorgproces	De fasen in het zorgproces die van belang zijn volgen uit de sectorspecifieke kwaliteitsvisie en dienen per fase te worden beschreven.

[21] Of *gedurende*, in het geval van een chronische behandeling.

1. [titel indicator]	
Methodologische criteria (ex ante beoordeling)	
Inhoudsvaliditeit	[Geef aan hoe inhoudsvaliditeit is beoordeeld, en wat de conclusie is voor de betreffende indicator. Dit conform indicatorstandaard, refereert aan *Bron* in eerste onderdeel van dit schema] Toelichting: De relatie tussen de geleverde zorg (of het ontbreken daarvan) en de zorguitkomsten is duidelijk. Voor *uitkomstindicatoren* betekent dit dat is aangetoond dat de gemeten uitkomst beïnvloedbaar is door de zorgaanbieder(s) waar de indicator betrekking op heeft. Voor *structuur- en procesindicatoren* betekent dit dat is aangetoond dat de gemeten structuur of processen ook daadwerkelijk de gewenste zorguitkomsten beïnvloeden. De inhoudsvaliditeit op setniveau is goed als de set van indicatoren de relevante aspecten van de geleverde zorg goed dekt. Hierbij wordt gelet op de relevante fasen (indicatie, proces van zorg zelf, uitkomsten) en de kwaliteitsdomeinen (effectiviteit, veiligheid, patiëntgerichtheid).
Populatievergelijkbaarheid	[Geef aan hoe populatievergelijkbaarheid is beoordeeld, en wat de conclusie is voor de betreffende indicator. Dit conform indicatorstandaard, refereert aan *Beschrijving populatieverschillen* in eerste onderdeel van dit schema] Toelichting: De indicatorwaarden van verschillende zorgaanbieders op een bepaalde indicator zijn vergelijkbaar. Dat wil zeggen dat de berekende indicatorwaarden daadwerkelijke verschillen in de kwaliteit van de geleverde zorg weerspiegelen en niet de verschillen in de patiëntenpopulaties van de zorgaanbieders.
Registratievergelijk-baarheid	[Geef aan hoe registratievergelijkbaarheid is beoordeeld, en wat de conclusie is voor de betreffende indicator. Dit conform indicatorstandaard] Toelichting: het proces van meten, registreren, aanleveren en verwerken van de voor de indicator benodigde gegevens dient juist (uniform en zonder meetfouten) te gebeuren.
Statistisch betrouwbaar onderscheiden	[Geef aan hoe statistisch betrouwbaar onderscheiden is beoordeeld, en wat de conclusie is voor de betreffende indicator. Dit conform indicatorstandaard] Een indicator dient het vermogen te hebben om zorgaanbieders met bovengemiddelde en ondergemiddelde indicatorwaarden te onderscheiden van gemiddeld scorende aanbieders. Bij dit criterium speelt het aantal waarnemingen dat een zorgaanbieder kan aanleveren een essentiële rol: bij een te laag aantal waarnemingen wordt de rol van toeval veelal te groot om betrouwbaar verschillen in prestaties te kunnen detecteren.
Opmerkingen	[beschrijving, bijv. beperkingen bij gebruik en interpretatie]

Bijlage 9: Ontwikkelagenda

Onderwerp	Onder-deel	Toelichting
1. Verken opname *alternatieve hierarchy of evidence*	H2	Zie bijvoorbeeld levels of evidence van NICE: http://www.nice.org.uk/niceMedia/pdf/GDM_Chapter7_0305.pdf
2. Ontwikkel een bijlage waarin de steekproeftrekking en correctie wordt beschreven	Nieuwe bijlage	Hoofdstuk 5 beschrijft de beoordeling van steekproeven. Een bijlage waarin geaccepteerde procedures worden beschreven kunnen een verantwoorde steekproeftrekking en correctie ondersteunen.
3. Ontwikkel bindende voorschriften voor het ontwikkelen en onderhouden van indicatoren	Nieuwe bijlage	Op dit moment bevat de standaard nog weinig bindende voor-schriften voor het ontwikkelen en onderhouden van indicatoren(sets). Een uitzondering zijn de voorschriften op het gebied van statistisch betrouwbaar onderscheiden. De ontwikkeling van bindende voorschriften vereist een procedure met consensus-vorming. Het is op dit moment niet duidelijk hoe succesvol een dergelijke procedure kan zijn.
4. Bundel adviezen (do's and don'ts) voor het ontwikkelen en onderhouden van indicatoren.	Nieuwe bijlage	Bundel adviezen voor het ontwikkelen en onderhouden van indicatoren(sets) zodat zij beter voldoen aan de toetscriteria in deze indicatorstandaard. Daarbij kan worden voortgebouwd op AIRE 2.0.