



Betrouwbaar onderscheiden

Een achtergrondstudie naar de statistische betrouwbaarheid en steekproefomvang bij het vergelijken van zorgaanbieders
september 2009

*Achtergrondstudie uitgevoerd door Xander Koolman (SiRM),
in opdracht van het programma Zichtbare Zorg*

Inhoudsopgave

1.	Inleiding.....	5
2.	Statistische betrouwbaarheid.....	7
3.	Onderscheidingsvermogen.....	8
4.	Uitkomstmaat en schaal.....	10
5.	Steekproefomvang per zorgaanbieder.....	12
6.	Samenvoegen van metingen.....	14
6.1	Indicatoren samenvoegen.....	14
6.2	Gebruik van voor- en nametingen.....	15
7.	Classificeren: significantie of relevante verschillen.....	16
8.	Validiteit en statistische betrouwbaarheid.....	17
9.	Afwegingen maken in de praktijk.....	18
10.	Conclusies.....	20
11.	Aanbevelingen.....	21
12.	Referenties.....	22
Bijlage I	De waarde van statistische betrouwbaarheid.....	23
Bijlage II	Berekening van statistische betrouwbaarheid.....	24
Bijlage III	Poweranalyse software.....	25

1. Inleiding

De kwaliteit van de zorg die zorgaanbieders bieden wordt gemeten zodat zorgconsumenten en zorginkopers die informatie kunnen gebruiken om zorgaanbieders te selecteren en te contracteren. Informatie over kwaliteitsverschillen dient daarom valide en betrouwbaar te zijn. De validiteit kan verstoord worden door meetfouten, case-mixverschillen en selectieve steekproeven. De statistische betrouwbaarheid wordt bepaald door de invloed van toeval op de metingen.

Toeval is het gevolg van de toevallige steekproef van patiënten, het toevallige meetmoment, en de toevallige gezondheid van de patiënten die in behandeling zijn. Toeval kan ervoor zorgen dat kwalitatief goede zorgaanbieders op het moment van de meting toch slecht scoren. We spreken van statistisch onbetrouwbare metingen wanneer een groot deel van de uitkomsten het gevolg zijn van toeval en niet van verschillen in kwaliteit. De invloed van toeval kan worden beperkt, bijvoorbeeld door het aantal metingen te verhogen.

Het verschil tussen validiteit en statistische betrouwbaarheid wordt duidelijk met behulp van een illustratie. Stel dat ziekenhuis A de noemer van een sterfteratio baseert op alle patiënten en dat ziekenhuis B de noemer baseert op enkel de patiënten die risico lopen. Wanneer beide ziekenhuizen evenveel sterfgevallen hebben op een even grote populatie (alle patiënten), dan scoort ziekenhuis A door de keuze van de noemer beter scoren B. In dat geval is de meting niet *valide*, omdat de indicatoren geen reëel beeld geven van de kwaliteitsverschillen per ziekenhuis [1,2]. Wanneer bij herhaling van de meting blijkt dat de scores zeer consistent zijn, dan zijn deze scores wel *statistisch betrouwbaar* [3,4,5]. Het is dus mogelijk dat een niet valide meting wel statistisch betrouwbaar is, en andersom. Dit rapport gaat enkel over de wijze waarop de invloed van toeval op onderzoek naar zorgkwaliteit verminderd kan worden.

Kwaliteitsscores die sterk door toeval worden beïnvloed zijn lastig te interpreteren door de gebruikers van die informatie. Een groter deel van de gebruikers zal beslissen om de informatie niet te gebruiken. Onbegrip over resultaten die sterk fluctueren door toeval kan eveneens leiden tot een lagere motivatie onder zorgaanbieders om aan de metingen mee te werken. In andere woorden, informatie over kwaliteit is waardevoller naarmate deze statistisch betrouwbaarder is.

Een voorbeeld om de invloed van toeval op kwaliteitsmetingen duidelijk te maken. Stel dat gemiddeld 2% van de verpleeghuispatiënten in Nederland decubitus ontwikkelt. Zorgaanbieder A weet dit percentage te reduceren tot 1%. De kwaliteitsmeting voor A wordt gebaseerd op 25 patiënten. Naar verwachting ontwikkelen 0,25 van de 25 patiënten decubitus. Tijdens een willekeurig gekozen meetmoment zal A veelal geen patiënten met decubitus hebben. Daardoor zal de kwaliteit van A als goed beoordeeld worden. Soms zal A echter wel een patiënt met decubitus bij de meting. In dat geval presteert A twee keer zo slecht ($1/25 = 4\%$) als het landelijk gemiddelde. De prestatie van A wisselt sterk als gevolg van het *toevallige* meetmoment.

Toeval kan verschillende effecten hebben. Zo zullen indicatoren die gevoelig zijn voor toeval sterk verschillen tussen de meetmomenten, zelfs indien de kwaliteit van zorg gelijk blijft. Door toeval is het mogelijk dat zorgaanbieders die over de gehele linie goede kwaliteit bieden, slecht zullen scoren op enkele indicatoren.

De statistische betrouwbaarheid van de huidige kwaliteitsinformatie is veelal beperkt. Bij de huidige ontwikkeling en analyse van kwaliteitsonderzoek ligt de nadruk op validiteit. Vrijwel elke keuze om de validiteit te verbeteren heeft echter invloed op de uiteindelijke statistische betrouwbaarheid. Het is belangrijk om ook expliciet rekening te houden met de statistische betrouwbaarheid in de ontwikkelingsfase van kwaliteitsonderzoek. De statistische betrouwbaarheid van huidige kwaliteitsinformatie is in veel gevallen eenvoudig te verbeteren.

Dit rapport reikt kennis aan waarmee statistische betrouwbaarheid van kwaliteitsinformatie kan worden verbeterd. Centraal staan het ontstaan en herkennen van toeval in het meten van kwaliteit en de technieken om de toevalsfactor te beheersen. Daarbij wordt onder andere ingegaan op de invloed van het aantal observaties of steekproefgrootte.

Het rapport is als volgt opgebouwd. Eerst wordt nader ingegaan op het begrip statistische betrouwbaarheid (hoofdstuk 2). Deze statistische betrouwbaarheid hangt sterk samen met het onderscheidingsvermogen (hoofdstuk 3), de gekozen uitkomstmaat en het gekozen afkappunt (hoofdstuk 4). Uiteindelijk bepalen de keuzen ten aanzien van de statistische betrouwbaarheid, het onderscheidingsvermogen en het afkappunt op de uitkomstschaal de vereiste steekproefomvang (hoofdstuk 5). Soms kan de steekproefomvang verkleind worden door metingen van verschillende meetmomenten of zorgaanbieders samen te voegen (hoofdstuk 6). Vervolgens wordt het classificeren van kwaliteitsinformatie besproken (hoofdstuk 7), alsmede de invloed van validiteit op statistische betrouwbaarheid (hoofdstuk 8). In hoofdstuk 9 volgen adviezen over het maken van afwegingen in de praktijk. Dit advies wordt afgerond met conclusies en de aanbevelingen in de laatste twee hoofdstukken. In de bijlagen wordt op enkele technische onderwerpen dieper ingegaan.

2. Statistische betrouwbaarheid

Hoewel statistische betrouwbaarheid veel grijswaarden kent, spreekt men in de praktijk toch van *of* betrouwbare *of* onbetrouwbare uitkomsten. In statistisch onderzoek is het gebruikelijk om te spreken van een betrouwbaar verschil wanneer de afwijking zo groot is dat deze *waarschijnlijk* niet door toeval wordt veroorzaakt. Wanneer de invloed van de toevalsfactor groot is, wordt de meting statistisch onbetrouwbaar genoemd. Een statistisch betrouwbare meting zal bij een herhaling een sterk vergelijkbare score opleveren (zie bijlage 1).

Significantie

Een betrouwbaar verschil zal *waarschijnlijk* niet door toeval worden veroorzaakt. Waarschijnlijk betekent hier dat louter toeval minder dan eens in de twintig keer een vergelijkbaar groot verschil zal opleveren [4,5]. Met andere woorden, een zorgaanbieder die niet werkelijk beter of slechter is maar enkel door toeval afwijkt, heeft slechts een kans van 5% (1 op 20) om ten onrechte als (significant) beter of slechter dan de gemiddelde zorgaanbieder te worden aangemerkt. *In statistische zin is het statistische betrouwbaarheidsniveau gedefinieerd als 1 minus het significantieniveau.* Bij een significantieniveau van 5% is het statistische betrouwbaarheidsniveau dus 95%. In de wetenschap is het gebruikelijk om te kiezen voor een significantieniveau van 5%. Bij wijze van uitzondering worden ook 10% (zwak significant) en 1% (sterk significant) gebruikt.

Bij een significantieniveau van 5% zal naar verwachting 5% van de gemiddeld presterende zorgaanbieders geduid worden als significant beter of slechter dan gemiddeld. De kans dat een zorgaanbieder voor de gehele set indicatoren door toeval wordt benadeeld (of bevoordeeld) is echter klein. Door het significantieniveau te verlagen kan het aandeel verkeerd toegekende prestaties van gemiddeld presterende zorgaanbieders worden verlaagd. Dit gaat echter ten koste van het onderscheidingsvermogen. Het onderscheidingsvermogen wordt uitgelegd in hoofdstuk 3.

3. Onderscheidingsvermogen

Statistische betrouwbaarheid is de kans dat een *gemiddelde* zorgaanbieder als gemiddeld wordt geduid. Onderscheidingsvermogen of *power* is de kans dat een *onderscheidende* zorgaanbieder onderscheiden wordt. Anders geformuleerd, onderscheidingsvermogen is de kans dat een werkelijk betere of slechtere zorgaanbieder ook als significant beter of slechter wordt geduid.

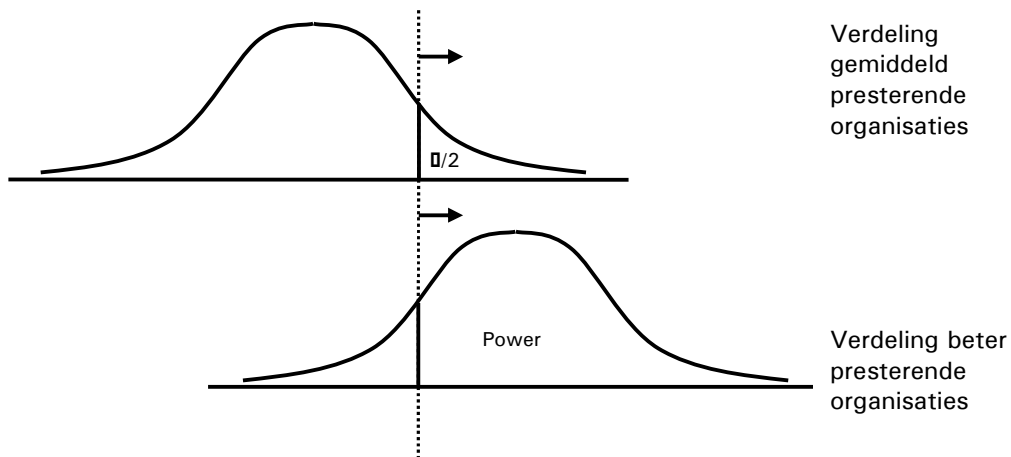
Net zoals het mogelijk is dat een verschil door toeval als significant wordt geduid, is het ook mogelijk dat een werkelijk verschil door toeval niet als significant wordt geduid. In de wetenschap is het gebruikelijk om te kiezen voor een onderscheidingsvermogen van minimaal 80% [5]. Dat impliceert dat organisaties die wezenlijk betere of slechtere zorg leveren dan gemiddeld, toch in 20% van de gevallen als gemiddeld worden geduid. Gebruikers van de informatie weten daardoor niet of zorgaanbieders die gemiddeld scoren in werkelijkheid toch slechte kwaliteit leveren. Hierdoor vermindert de gebruikswaarde van de kwaliteitsscores .

De uitruil tussen statistische betrouwbaarheid en onderscheidingsvermogen

Zoals in hoofdstuk 2 aangegeven bestaat er een uitruil tussen statistische betrouwbaarheid en onderscheidingsvermogen. Deze wordt verduidelijkt met figuur 1. Daarin staan twee normale verdelingen van prestaties. De eerste verdeling beschrijft de verdeling van de geschatte prestaties bij zorgaanbieders die alle gemiddelde kwaliteit van zorg leveren. De prestaties wijken enkel door toeval af van het gemiddelde. Het is onwaarschijnlijk dat de zorgaanbieders bij een meting precies de gemiddelde kwaliteit scoren. Een kleine afwijking van het gemiddelde zal dus veel voorkomen. Een grote afwijking zal minder vaak voorkomen. Dit is de basis van de populaire 'normale verdeling' die klokvormig van aard is en dunne uiteinden of staarten heeft. De tweede verdeling in figuur 1 beschrijft de verdeling van zorgaanbieders die zorg van een beter dan gemiddelde kwaliteit leveren. Ook bij deze aanbieders speelt toeval in de meting van hun prestaties een rol, waardoor het in de praktijk mogelijk is dat een betere aanbieder slechter scoort dan een gemiddelde aanbieder.

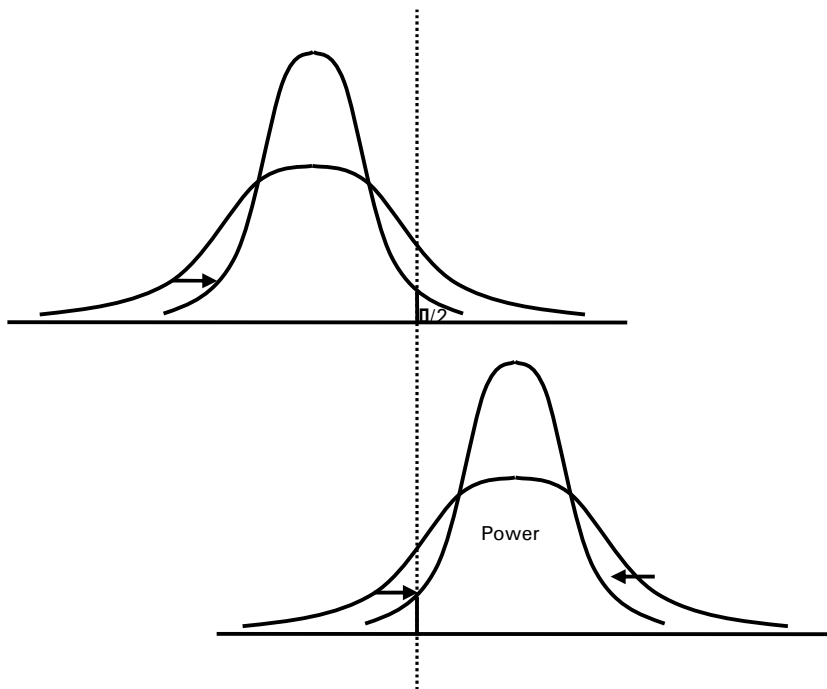
Zoals hierboven aangegeven kan de statistische betrouwbaarheid worden vergroot door het significantieniveau te verlagen. In de figuur komt dit overeen met het naar rechts schuiven van de verticale lijn. Hierdoor wordt de oppervlakte $\alpha/2$ (helft significantieniveau) kleiner. De figuur illustreert dat indien het significantieniveau kleiner wordt, het onderscheidingsvermogen (oppervlakte onder de curve 'power') eveneens kleiner wordt. Een grotere statistische betrouwbaarheid gaat dus gepaard met een lager onderscheidingsvermogen.

Figuur 1 Uitrui tussen statistische betrouwbaarheid en onderscheidingsvermogen



Toch is het mogelijk om zowel statistische betrouwbaarheid als onderscheidingsvermogen te verbeteren. Dit kan door de invloed van toeval op de uitkomsten te verkleinen. Figuur 2 laat zien hoe zowel de statistische betrouwbaarheid als het onderscheidingsvermogen verbeteren wanneer de prestaties preciezer worden geschat: de oppervlakte ' $\alpha/2$ ' neemt af en de oppervlakte 'power' neemt toe.

Figuur 2 Verbeteren van statistische betrouwbaarheid en onderscheidingsvermogen



Preciezer schatten kan door een meting op een groter aantal observaties (waarnemingen) te baseren, maar een andere aanpak is ook mogelijk. De volgende hoofdstukken beschrijven de belangrijkste opties.

4. Uitkomstmaat en schaal

Indicatoren worden veelal onderscheiden in structuur-, proces- en uitkomstindicatoren. Structuur- en procesindicatoren worden veelal niet op patiëntniveau gemeten. De kwaliteit gemeten op basis van deze indicatoren is hierdoor niet afhankelijk voor de invloed van de (toevallige) samenstelling van de patiënten die aan de meting deelnemen. Echter, voor een goed en compleet beeld van kwaliteit van zorg zijn ook uitkomstindicatoren nodig[6].

Er zijn algemene en ziektespecifieke uitkomstindicatoren. De keuze voor een algemene of ziektespecifieke uitkomstindicator beïnvloedt de uitkomstschaal waarop de indicator wordt gemeten. Deze keuze heeft daarmee invloed op de statistische betrouwbaarheid en het onderscheidingsvermogen. Dit hoofdstuk gaat over keuze van de uitkomstmaat en schaal.

Algemene versus ziektespecifieke uitkomstindicatoren

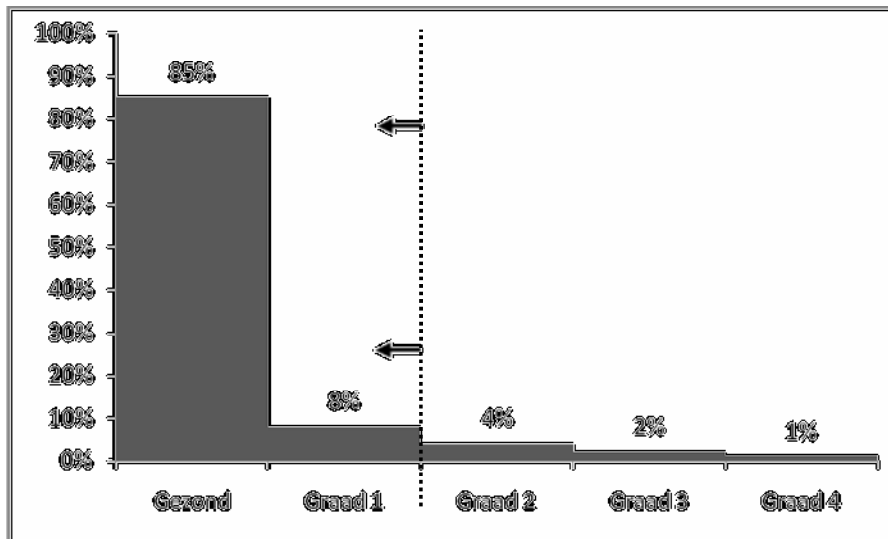
De keuze voor algemene of ziektespecifieke uitkomstindicatoren is van invloed op de statistische betrouwbaarheid van de kwaliteitsmeting. Algemene uitkomstmaten meten de totale gezondheidstoestand, en ziektespecifieke uitkomstmaten meten uitkomsten die direct door de ziekte worden beïnvloed. Deze laatste groep is in de regel gevoeliger voor kleine veranderingen en vereist daarom kleinere aantallen observaties dan algemene uitkomstindicatoren. Zoals de naam al aangeeft zijn ziektespecifieke indicatoren echter niet te gebruiken voor patiënten met verschillende ziekten. Daar ligt de kracht van de algemene uitkomstmaten die gebruikt kunnen worden voor uiteenlopende groepen patiënten. Algemene uitkomstmaten zijn daarmee aantrekkelijk voor prestatiemeting van zorgaanbieders met relatief veel, maar sterk verschillende patiënten.

Keuze van afkappunt

De uitkomst van een indicator kan worden vastgelegd als een dichotome (0 of 1) variabele. Voor een indicator als sterfte is dat voor de hand liggend. Voor veel van de gebruikte indicatoren is dat echter niet noodzakelijk. In die gevallen kent de uitkomst eigenlijk vele grijswaarden, maar deze worden dan gereduceerd tot twee uitkomsten: goed of slecht. De keuze van het afkappunt is mede bepalend voor de verhouding tussen de uitkomsten.

De samenhang tussen afkappunt en statistische betrouwbaarheid wordt geïllustreerd met de indicator decubitus. Decubitus wordt veelal onderverdeeld in vier gradaties, waarbij graad 4 de meest ernstige vorm is. Figuur 3 geeft een hypothetische verdeling van het voorkomen van de verschillende graden decubitus. Om deze schaal van 0 tot 4 te transformeren tot een dichotome indicator, wordt de grens veelal gelegd bij de tweede graad. Deze grens wordt in de figuur weergegeven met behulp van een stippellijn. Bij dit afkappunt komt decubitus voor bij 7% van een patiëntengroep. In dat geval zijn er 382 observaties per zorgaanbieder nodig om aan te tonen dat een zorgaanbieder met de helft van het landelijke gemiddelde significant beter presteert. Wanneer het afkappunt verschuift naar graad 1 komt decubitus voor bij 15% van een patiëntengroep. Dan daalt het aantal benodigde observaties naar 166. Indien het afkappunt naar rechts wordt opgeschoven zodat decubitus enkel meetelt vanaf de derde graad, zijn 921 waarnemingen nodig. De keuze van afkappunt heeft dus bij deze indicator een grote invloed op de statistische betrouwbaarheid van de scores. Zie tabel 1 voor meer voorbeelden over de invloed van afkappunten.

Figuur 3 Voorkomen van decubitus naar graad van ernst



Ongeacht de keuze van afkappunt zal door dichotomiseren altijd informatie verloren gaan. Wanneer informatie verloren gaat neemt de statistische betrouwbaarheid in de regel af. Een ordinale schaal is een schaal waarbij de uitkomsten wel geordend kunnen worden, maar het niet duidelijk is of de verschillen tussen de uitkomsten een vergelijkbare afstand hebben (zoals graad van decubitus). Bij een intervallschaal is dat het geval (zoals verondersteld wordt bij de CQ-index).

Het verlies aan statistische betrouwbaarheid door dichotomisatie hangt sterk samen met de spreiding in de gemeten uitkomst en met de relatie tussen de zorgaanbieders en de extreme uitkomsten. Zo zal zonder dichotomisatie een zorgaanbieder met extreme prestaties eerder significant zijn, dan een zorgaanbieder met gematigde prestaties. Uitgaande van de in figuur 3 getoonde verdeling van het voorkomen van decubitus naar graad van ernst zijn bij deze intervallschaal voor decubitus slechts 60 observaties nodig, wezenlijk minder dan de 166 voor het laagste afkappunt.

De keuze tussen dichotome, ordinale of intervallschaal-uitkomsten is niet noodzakelijk een keuze tussen statistische betrouwbaarheid en validiteit. In veel gevallen is de keuze van het afkappunt arbitrair en zijn de uitkomsten gevoelig voor die toevallige keuze. Hierdoor komt ook de validiteit van de meting in gevaar. Daarnaast heeft dichotomisatie tot gevolg dat extreme uitkomsten en gematigde uitkomsten gelijk behandeld worden. Dit is onwenselijk omdat het een zorgaanbieder demotiveert om goede zorg te leveren aan een cliënt die ongeacht de kwaliteit van de behandeling boven of onder het afkappunt terecht komt. In het volgende hoofdstuk wordt de invloed van de statistische betrouwbaarheid, het onderscheidingsvermogen en het afkappunt op de steekproefomvang verduidelijkt.

5. Steekproefomvang per zorgaanbieder

De omvang van de steekproef is de meest voor de hand liggende factor om de statistische betrouwbaarheid te beïnvloeden. De steekproefomvang per zorgaanbieder is het aantal observaties per zorgaanbieder waarop een kwaliteitsmeting is gebaseerd. Meestal bestaat één observatie uit de meting van een uitkomst bij één patiënt. In de praktijk staat een meting veelal gelijk aan een kwaliteitsindicator per zorgaanbieder. Een groter aantal observaties leidt tot een preciezere en daardoor statistisch meer betrouwbare meting.

De relatie tussen statistische betrouwbaarheid, onderscheidingsvermogen en afkappunt kan worden verduidelijkt aan de hand van een rekenvoorbeeld. In dat rekenvoorbeeld is het aantal benodigde observaties per zorgaanbieder afhankelijk van de hoogte van het significantieniveau, het onderscheidingsvermogen en de keuze van het afkappunt.

Een rekenvoorbeeld

Tabel 1 laat zien hoeveel observaties nodig zijn om de kwaliteit van een zorgaanbieder die het voorkomen van decubitus tot de helft van het landelijke gemiddelde heeft weten te reduceren, als significant beter te duiden. De aantallen zijn afhankelijk van het significantieniveau (α) het onderscheidingsvermogen (power) en het aandeel van de populatie dat decubitus heeft. Voor een leesvoorbeeld gaan we uit van een significantieniveau van 5% (betrouwbaarheid van 95%) en een onderscheidingsvermogen van 80%. Bij deze waarden blijkt uit tabel 1A dat een indicator met gemiddeld 10% positief scorende patiënten, 260 observaties per zorgaanbieder vereist om de kwaliteit van een zorgaanbieder met de helft minder dan gemiddeld positieve scores als significant beter dan gemiddeld te duiden. Indien het significantieniveau wordt verlaagd naar 0,001% dan is een steekproef van 573 observaties per zorgaanbieder vereist.

Tabel 1A Benodigd aantal waarnemingen bij een gemiddelde van 10%

		Power				
		0,5	0,8	0,9	0,95	0,99
α	0,1	118	205	262	314	426
	0,05	159	260	324	383	506
	0,01	260	389	467	537	683
	0,001	412	573	668	751	923

Tabel 1B Benodigd aantal waarnemingen bij een gemiddelde van 25%

		Power				
		0,5	0,8	0,9	0,95	0,99
α	0,1	41	72	92	111	151
	0,05	55	90	113	134	178
	0,01	89	134	162	186	238
	0,001	139	196	229	259	320

Tabel 1C Benodigd aantal waarnemingen bij een gemiddelde van 50%

		Power				
		0,5	0,8	0,9	0,95	0,99
α	0,1	15	27	35	43	59
	0,05	20	34	43	51	68
	0,01	31	49	59	69	90
	0,001	48	70	83	94	118

Het benodigd aantal waarnemingen hangt samen met het aandeel patiënten dat positief scoort op de uitkomst. Daarom is tabel 1 onderverdeeld in drie delen (A, B en C) die elk de benodigde aantallen waarnemingen presenteren maar voor afwijkende percentages positief scorende patiënten (10, 25 en 50%). Uitgangspunt van elke berekening is dat de kwaliteit van een zorgaanbieder met de helft van de gemiddelde positieve score kan worden geduid als significant beter. Omdat het in de wetenschappelijke literatuur gebruikelijk is om een significantieniveau van 0,05 en een power van 0,8 te kiezen, is het aantal observaties bij die waarden vet weergegeven.

Zoals blijkt uit de tabellen, vereisen dichotome indicatoren waarbij beide uitkomsten sterk ongelijk verdeeld zijn, meestal veel observaties. Het kan zijn dat het aantal vereiste observaties groter is dan het aantal patiënten dat behandeld wordt. In dat geval biedt het vergroten van de steekproefomvang geen soelaas. Het samenvoegen van metingen kan dan uitkomst bieden. Hierover gaat het volgende hoofdstuk.

6. Samenvoegen van metingen

De statistische betrouwbaarheid kan worden verbeterd door metingen samen te voegen. Zo kunnen indicatoren, patiëntgroepen of meetmomenten per zorgaanbieder worden samengevoegd. Samenvoegen van metingen dient echter met beleid te gebeuren. Naast samenvoegen kan de statistische betrouwbaarheid worden verhoogd met behulp van voor- en nametingen. Deze benaderingen worden in dit hoofdstuk besproken.

6.1 Indicatoren samenvoegen

Wanneer de informatie op indicatorniveau onvoldoende betrouwbaar is, dan kunnen metingen van verschillende indicatoren worden samengevoegd. Zo kunnen bijvoorbeeld alle indicatoren van een zorgaanbieder worden samengevoegd. Vervolgens zou dan kunnen worden berekend of de zorgaanbieder als geheel significant afwijkt van het gemiddelde. Statistisch significante uitspraken op indicatorniveau zijn dan echter niet mogelijk.

Deze aanpak is vooral interessant wanneer de prestaties 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 de statistische betrouwbaarheid van een prestatieoordeel zal toenemen. Wanneer prestaties voor indicatoren niet of nauwelijks samenhangen zullen de indicatoren waarop slecht wordt gescoord, worden gecompenseerd door indicatoren waarop goed wordt gescoord. In dat geval zal de statistische betrouwbaarheid weinig toenemen en mogelijk zelfs afnemen.

Bij een lage correlatie tussen prestaties op verschillende indicatoren is het aan te raden enkel die indicatoren te combineren die wel met elkaar correleren. Selecties kunnen gemaakt worden op basis van statistische en inhoudelijke argumenten. Zo kan op basis van statistisch onderzoek besloten worden welke prestaties sterk met elkaar correleren. Dit heeft wel tot nadeel dat de samengevoegde indicatoren waarschijnlijk niet meer te interpreteren zijn. 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 verminderde infecties bij verschillende behandelingen, of kan worden verwacht dat een tekort aan personeel in een verpleeghuis leidt tot meer fixatie en meer decubitus.

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 statistische betrouwbaarheid voldoen. Wanneer in de praktijk blijkt dat deze samengestelde indicatoren met elkaar correleren dan leidt samenvoeging tot een verhoogde statistische betrouwbaarheid van een interpreteerbare maat. Een voorbeeld van een samengestelde indicator is de CQ-index waarvoor Likert-schalen worden geconstrueerd op basis van met elkaar samenhangende items.

Patiëntgroepen samenvoegen

Het samenvoegen van cliëntgroepen gaat gepaard met vergelijkbare overwegingen. Zo is het mogelijk dat de prestaties van een zorgaanbieder voor verschillende patiëntgroepen sterk samenhangen. In dat geval levert het samenvoegen veel statistische betrouwbaarheidswinst op. Als gevolg van de samenvoeging is het dan mogelijk betrouwbare uitspraken over de gemiddelde kwaliteit van een zorgaanbieder te doen. Uitspraken per patiëntgroep zijn echter niet meer mogelijk. Wanneer de prestaties tussen patiëntgroepen niet of nauwelijks samenhangen, levert het samenvoegen weinig extra statistische betrouwbaarheid op.

Meetmomenten samenvoegen

Naast het samenvoegen van indicatoren en patiëntgroepen is het mogelijk om herhaalde metingen over een langere tijd samen te voegen. Vervolgens kan de significantie worden bepaald op basis van de gemiddelde prestatie van deze verschillende meetmomenten. De statistische betrouwbaarheidswinst zal groot zijn omdat deze prestaties waarschijnlijk sterk met elkaar samenhangen. Daarnaast blijven de resultaten goed te interpreteren. Nadeel is wel dat een verandering in kwaliteit minder snel zichtbaar wordt omdat de scores van eerdere metingen dit effect dempen.

Benodigde aantallen bij samenvoeging van metingen

Wanneer de prestaties volkomen met elkaar samenhangen (100%) dan is het aantal benodigde observaties per meting het totaal aantal observaties gedeeld door het aantal metingen. Dus bij een enkele meting 260 (zie tabel 1A) en bij twee metingen 130. Indien de prestaties tussen de metingen echter geheel niet samenhangen dan gaat het vereiste aantal waarnemingen per meetmoment juist omhoog tot 550. Eventuele voordelen van samenvoeging kunnen afnemen wanneer de waarnemingen zelf samenhangen. Dit kan bijvoorbeeld wanneer de metingen betrekking hebben op dezelfde patiënt. In dat geval is het veelal aantrekkelijk om de meting te baseren op voor- en nametingen op patiëntniveau.

6.2 Gebruik van voor- en nametingen

De statistische betrouwbaarheid van een meting verbetert indien de toevalsvariatie kan worden gereduceerd. 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 bij een gelijk aantal respondenten 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. Tabel 2 laat zien hoe de benodigde aantallen respondenten afnemen wanneer de correlatie tussen de voor- en nameting (ρ) toeneemt. Het verrichten van een voor- en nameting kan grote invloed hebben op de statistische betrouwbaarheid.

Tabel 2 Benodigd aantal respondenten bij voor- en nametingen

		Aantal metingen		
		2	3	4
Rho	0	260	195	173
	0,2	208	156	139
	0,4	156	117	104
	0,6	104	78	70
	0,8	52	39	35

7. Classificeren: significantie of relevante verschillen

Bij de presentatie op bijvoorbeeld KiesBeter.nl of andere consumentensites wordt kwaliteit veelal vereenvoudigd en vergelijkbaar weergegeven door de kwaliteitsscores te classificeren in drie of vijf klassen. Zo presenteert kiesBeter.nl prestaties voor de verpleeg- en verzorgingshuizen in vijf klassen. Daarbij krijgt een verpleeghuis dat tot de slechtste klasse behoort één ster en vijf sterren wanneer het tot de beste klasse behoort. Voor de CQ-index in de verpleeg- en verzorgingssector zijn de prestaties ingedeeld in vijf klassen. De middelste categorie met drie sterren bevat de zorgaanbieders die niet significant afwijken van het gemiddelde. Vier sterren krijgen zorgaanbieders die significant beter presteren dan gemiddeld, maar niet significant beter dan de gemiddelde bovengrens van de statistische betrouwbaarheidsintervallen van alle zorgaanbieders. Organisaties die wel significant beter presteren dan de gemiddelde bovengrens krijgen vijf sterren.

Deze CQ-index sterrenindeling is daarmee enkel gebaseerd op statistische significantie. De vraag of de verschillen ook voldoende groot zijn om *relevant* te zijn doet er voor de sterrenindeling niet toe. Een verschil is relevant als het als belangrijk wordt gezien door de gebruiker van de informatie. Het is mogelijk dat verschillen die als relevant of wezenlijk worden ervaren niet statistisch significant zijn. Andersom is het mogelijk dat significante verschillen voor de informatiegebruiker niet wezenlijk zijn. Stel bijvoorbeeld dat een meting gebaseerd is op twintig waarnemingen voor elke aanbieder, en dat bij aanbieder A vijf patiënten decubitus hebben ontwikkeld en bij aanbieder B geen enkele. Het verschil van 25% decubitus is relevant omdat het wordt gezien als een belangrijk verschil. Toch zal dat verschil waarschijnlijk niet statistisch significant zijn omdat de studie door het lage aantal waarnemingen onvoldoende onderscheidingsvermogen heeft. Evenzeer is het mogelijk dat een studie met een zeer groot aantal observaties aantoont dat cliënten van grote verpleeghuizen gemiddeld 0,05% minder decubitus ontwikkelen en dat het verschil met kleine verpleeghuizen statistisch significant is. Is dit significante verschil ook relevant of betekenisvol voor de keuze van zorgaanbieder? Dat zal de consument moeten bepalen, en om dat te kunnen dient hij het verschil tussen significantie en relevantie op waarde te schatten. Dat is voor de gemiddelde consument lastig indien beide gegevens apart worden weergegeven, en onmogelijk wanneer de prestatie-informatie enkel in de vorm van sterren wordt gepresenteerd. Daarom kan het handig zijn het aantal waarnemingen zo te kiezen dat relevante verschillen significant zullen zijn en irrelevante verschillen niet. In de praktijk zal het aantal waarnemingen echter niet per indicator worden vastgesteld, waardoor het combineren van significantie en relevantie in een score lastig zal blijken.

Er kleven meer nadelen aan een classificatie met vijf sterren. Indien het aantal waarnemingen beperkt is zullen één ster en vijf sterren in de praktijk niet worden toegekend. Het risico is dat consumenten het verschil tussen een goede en een slechte organisatie als klein interpreteren, terwijl het werkelijke verschil groot kan zijn. Indien het wenselijk is dat een bepaald deel van de aanbieders één ster of vijf sterren scoren, dan is een wezenlijk groter aantal waarnemingen vereist. Het is aan te raden een classificatie met drie sterren te gebruiken. Het benodigde aantal waarnemingen zal niet dalen ten opzichte van een vijf sterren classificatie. Een goede inschatting van de extra benodigde waarnemingen is te maken met behulp van simulatietechnieken. Deze techniek is voor deze vraag niet toegepast.

8. Validiteit en statistische betrouwbaarheid

De validiteit van informatie over kwaliteitsverschillen hangt onder ander af van een goede correctie voor verschillen in case-mix tussen zorgaanbieders. Verschillen in case-mix kunnen in sommige situaties worden gecorrigeerd door exclusie van bepaalde patiënten voor de metingen (schonen). In andere situaties is (indirecte) standaardisatie vereist. Standaardisatie is correctie voor verschillen in case-mix tussen zorgaanbieders met behulp van statistische analyse. Daarbij is de meest gebruikte techniek het vergelijken van de waargenomen uitkomsten met de verwachte uitkomsten op basis van de case-mix. Bij decubitus is in de verpleeg- en verzorgingssector mede gecorrigeerd voor case-mix door cliënten die met decubitus in zorg werden opgenomen niet bij de berekening van de indicator mee te nemen (schoning). Hierdoor neemt het aantal patiënten met decubitus dat meetelt wezenlijk af. Bijgevolg neemt ook de statistische betrouwbaarheid af. Indien de case-mix correctie was uitgevoerd door te standaardiseren dan was de statistische betrouwbaarheid waarschijnlijk groter geweest. Naast dit voordeel heeft standaardisatie ook tot voordeel dat de inclusie van cliënten met decubitus ten tijden van de aanvang ook een prikkel introduceert om goede zorg te leveren aan deze cliënten.

Hoewel standaardisatie betrouwbaarder resultaten op kan leveren dan schoning, leidt ook deze vorm van correctie tot een verlaging van de statistische betrouwbaarheid ten opzichte van niet corrigeren. De mate waarin de statistische betrouwbaarheid afneemt is afhankelijk van de samenhang tussen de prestaties en de case-mix variabele. De benodigde hoeveelheid waarnemingen verdubbelt wanneer de samenhang 50% is. In de praktijk is de samenhang waarschijnlijk minder groot, omdat het op voorhand niet waarschijnlijk is dat de gestandaardiseerde prestaties sterk met de case-mix samenhangen. Wij verwachten daarom dat schonen in de regel tot statistisch minder betrouwbare uitkomsten leidt dan standaardisatie.

9. Afwegingen maken in de praktijk

In de praktijk zal de vereiste statistische betrouwbaarheid samenhangen met het doel waarvoor de resultaten gebruikt zullen worden. Zo is het waarschijnlijk dat zorgaanbieders en de Inspectie voor de gezondheidszorg voor kwaliteitscontroles een lager niveau van statistische betrouwbaarheid vereisen. Zij zijn immers in staat de informatie op waarde te schatten en, indien nodig, verder onderzoek te initiëren. Zorginkopers en zorgconsumenten vereisen waarschijnlijk betrouwbaarder informatie. Zorginkopers zijn afhankelijk van de juistheid van de informatie voor het afsluiten van (prestatie)contracten en de zorgconsument voor de selectie van een zorgaanbieder. Het is voor hen moeilijk om aan aanvullende informatie te komen.

Of, en zo ja welke, maatregelen nodig zijn om de statistische betrouwbaarheid te verbeteren, verschilt sterk van situatie tot situatie. Toch is het mogelijk om een ordening in de maatregelen aan te brengen die in veel gevallen geschikt zal zijn. Maatregelen die de statistische betrouwbaarheid én de validiteit verbeteren genieten de voorkeur. In de casus hieronder valt de keuze op het gebruik van voor- en nametingen. Indicatorscores op basis van voor- en nametingen zijn minder gevoelig voor verstoring door case-mixverschillen.

Vervolgens kan gekozen voor maatregelen die naast de statistische betrouwbaarheid eveneens de prikkelstructuur bevorderen. Hieronder valt de keuze voor standaardisatie in plaats van schoning omdat de extra geïncorporeerde patiënten een prikkel introduceren om ook voor deze patiënten kwalitatief goede zorg te verlenen. Ook de keuze om de indicator te meten met een gedetailleerde schaal in plaats van een dichotome (0 of 1) schaal verbetert de prikkels. Een gedetailleerde schaal stimuleert de zorgaanbieder om voor elke patiënt goede zorg te leveren, ook wanneer de patient door deze behandeling niet dusdanig veel verbeterd dat zijn score op een dichotome schaal verandert (score blijft aan zelfde zijde van het afkappunt).

Minder aantrekkelijk zijn de maatregelen die de informatiewaarde verminderen. Maatregelen als het verhogen van het significantieniveau en het verlagen van het onderscheidingsvermogen verminderen de waarde van de informatie voor de gebruikers en tasten de motivatie van de zorgaanbieders om te meten aan. Afwijkende significantieniveaus tussen indicatoren kunnen leiden tot verwarring bij de gebruikers van de informatie. Ook het samenvoegen van metingen kan een optie zijn wanneer de metingen sterk samenhangen.

Het is wellicht het minst aantrekkelijk om kwaliteitsmeting te baseren op structuur- en procesindicatoren in plaats van uitkomstindicatoren. Structuur- en procesindicatoren die worden gemeten op zorgaanbiederniveau zijn niet gevoelig voor toeval (wel van meetfout). Zelfs op patiëntniveau gemeten procesindicatoren minder gevoelig voor patiëntgerelateerde toevalsvariatie. Immers, procesindicatoren representeren behandelbeleid (wel of niet sederen), en dat is minder gevoelig voor toevalsvariatie. Dit maakt structuur en procesindicatoren aantrekkelijk. Maar, de uitkomst is juist van belang. En de relatie tussen structuur- en procesindicatoren aan de ene kant en uitkomsten aan de andere kant, is vaak zwak. Het te verwachten dat deze relatie nog verder zal afzwakken indien zorgaanbieders worden afgerekend op structuur- en procesindicatoren, en daardoor pervers gedrag gaan vertonen [7].

Een praktisch voorbeeld

Dit rapport licht toe welke factoren van invloed zijn op de statistische betrouwbaarheid van kwaliteitsonderzoek aan de hand van rekenvoorbeelden. In de praktijk zullen keuzen die daarin gemaakt worden samenhangen met elkaar en met factoren die niet beïnvloedbaar zijn. Daardoor kan bij voorbeeld de keuze voor een specifieke indicator voor een specifieke sector wezenlijk anders uitpakken in een andere sector. Hieronder wordt een rekenvoorbeeld gegeven om te illustreren hoe een aantal keuzen gezamenlijk de statistische betrouwbaarheid beïnvloeden. De onderliggende aannames voor het rekenvoorbeeld zijn niet getoetst en zullen in de praktijk afwijken van de werkelijke waarden.

Stel nu dat de kwaliteit van zorgaanbieders die een gering aantal patiënten gedurende langere tijd in behandeling hebben gemeten moet worden. Zorgaanbieders met een klein aantal patiënten, hebben veelal een klein aantal patiënten omdat zij deze patiënten intensief behandelen. Indien de behandeling kortdurend is, kan gekozen worden voor meerdere meetmomenten binnen een jaar om zo de vereiste aantallen waarnemingen te behalen. In een aantal gevallen is dat echter niet mogelijk omdat de patiënten niet alleen intensief maar ook langdurig worden behandeld. Voorbeelden van dergelijke zorg zijn te vinden in de gehandicaptenzorg en de geestelijke gezondheidszorg.

De statistische betrouwbaarheid kan niet worden vergroot door extra meetmomenten in een jaar te verrichten omdat door de langdurige relatie weinig nieuwe patiënten kunnen worden toegevoegd. Wel kunnen steeds meer metingen gedaan worden per patiënt naarmate de behandelrelatie langer duurt. Echter, als de oorzaak van langdurige behandelrelaties is dat de patiënten niet snel herstellen, dan zal er weinig variatie in de uitkomst van de de herhaalde metingen waarneembaar zijn. Daardoor zal het vereiste aantal waarnemingen hoog zijn. Daarom is het aan te raden naast voor- en nametingen de statistische betrouwbaarheid te optimaliseren. Zo is het mogelijk om van 260 waarnemingen naar bijvoorbeeld 104 waarnemingen te gaan met behulp van voor- en nametingen (zie tabel 2) en de 0-1 uitkomstschaal te vervangen door een interval schaal. Hierdoor kan het benodigde aantal waarnemingen verder dalen tot 26 (zie hoofdstuk 4). Verdere reductie kan worden bereikt door samenhangende indicatoren samen te voegen. Uitgaande van twee indicatoren met een correlatie van ongeveer 0.5 kan het benodigde aantal tot 18 verminderd worden. Tot slot kan het benodigd aantal waarnemingen verminderd worden door meerdere indicatoren samen te voegen.

Indien een kwaliteitsindicator onvoldoende statistisch betrouwbaar blijkt, kan gekozen worden om deze niet openbaar te maken. Deze informatie is echter toch waardevol. Immers, zorgaanbieder A uit het voorbeeld in de inleiding zal tijdens de meeste metingen goed scoren. Anders geformuleerd: gemiddeld zullen de goed scorende zorgaanbieders betere zorg leveren. En vice versa, een goede zorgaanbieder zal gemiddeld beter scoren over alle indicatoren of meetmomenten. Minder betrouwbare indicatoren hoeven daarom niet bij voorbaat genegeerd te worden. De presentatie van onbetrouwbare indicatoren aan zorgconsumenten vereist echter zorg.

10. Conclusies

Goede indicatoren meten wat ze beogen te meten (ze zijn valide) en zijn niet gevoelig voor toeval (ze zijn statistisch betrouwbaar). Die statistische betrouwbaarheid is van grote waarde voor de gebruikers van de informatie, bij huidige indicatoren veelal beperkt en in veel gevallen eenvoudig te verbeteren.

De statistische betrouwbaarheid is van belang omdat weinig betrouwbare indicatorscores sterk zullen fluctueren tussen metingen, waardoor de kwaliteit van een zorgaanbieder lastig te beoordelen is. Zorggebruikers zullen de informatie wantrouwen en zorginkopers zullen per abuis slechte zorgaanbieders contracteren. Door de zwakke relatie tussen de geleverde kwaliteit van zorg en de indicatorscores zullen zorgaanbieders gedemotiveerd raken om te meten *en* om te investeren in kwaliteit van zorg.

Met behulp van een aantal maatregelen kan de statistische betrouwbaarheid van bestaande en nieuw te ontwikkelen indicatoren wezenlijk worden verbeterd. Statistische betrouwbaarheid kan worden verbeterd met:

- gebruik van voor- en nametingen
- corrigeren voor case-mix door middel van standaardisatie in plaats van schonen
- indicatorregistratie op gedetailleerde in plaats van 0-1 schaal
- prestaties rapporteren in zo min mogelijk klassen (drie in plaats van vijf)
- samenvoeging van metingen en/ of indicatoren

De meeste van deze keuzen zijn niet direct strijdig met de validiteit van een indicator. Toch zal dat soms wel het geval zijn. Ook zullen sommige keuzen leiden tot extra kosten. Het is daarom aan te raden om vooraf in kaart te brengen wat de gevolgen zullen zijn van maatregelen die de statistische betrouwbaarheid verhogen. Dat is mogelijk met behulp van een analyse van het onderscheidingsvermogen van een indicator, beter bekend als poweranalyse (zie bijlage 3). In dit rapport is deze techniek toegepast om de rekenvoorbeelden te maken en zo inzicht te geven in de invloed van factoren die de statistische betrouwbaarheid bepalen. De werkelijke invloed van die factoren zal van indicator tot indicator en van situatie tot situatie variëren.

11. Aanbevelingen

De statistische betrouwbaarheid van kwaliteitsonderzoek hangt sterk samen met de methodologische keuzen die gemaakt worden tijdens het ontwikkelen van een indicator. Daarom is de eerste aanbeveling:

1. Maak de statistische betrouwbaarheid- en poweranalyse een vast onderdeel van indicatorontwikkeling.

Deze *a priori* poweranalyse geeft een beeld van de verwachte statistische betrouwbaarheid van het onderzoek. Indien deze lager is dan vereist, dan kan de indicator worden aangepast of in het meest ernstige geval worden geschrapt. Wanneer de statistische betrouwbaarheid keuzen vereist die strijdig zijn met andere eisen zoals validiteit en kosten, dan kan een poweranalyse de noodzakelijke informatie opleveren om een optimale uitruil te maken tussen al deze doelen.

Een poweranalyse vereist een groot aantal aannamen. Veel van deze aannamen kunnen vooraf worden getoetst met behulp van data uit (internationale) studies. Toch zal voor sommige aannamen geen onderbouwing kunnen worden gevonden of zal achteraf blijken dat die onderbouwing niet goed is. Daarom is de tweede aanbeveling:

2. Maak statistische betrouwbaarheid- en poweranalyse een vast onderdeel van indicatoronderhoud.

Deze *a posteriori* poweranalyse lijkt wellicht overbodig, omdat uit de meting direct blijkt of, en hoeveel, zorgaanbieders significant afwijken van het gemiddelde. Toch kan een achteraf analyse aanvullend inzicht geven over het onderscheidingsvermogen en de wijze waarop de statistische betrouwbaarheid verder kan worden verbeterd. Meer informatie over poweranalyses staat in bijlage 3.

De mate waarin statistische betrouwbaarheid wenselijk is hangt af van de meerkosten van een toename aan de ene kant en de meeropbrengsten aan de andere kant. Het is niet meer lonend om in extra statistische betrouwbaarheid te investeren wanneer de meerkosten hoger worden dan de meeropbrengsten. Daar waar de meerkosten in veel gevallen zeer tastbaar zijn geldt dat niet voor de meeropbrengsten. Dat zou gemakkelijk kunnen leiden tot een onderinvestering in de kwaliteit van kwaliteitsonderzoek. Daarom is de derde aanbeveling:

3. Onderzoek de waarde die gebruikers toekennen aan statistische betrouwbaarheid en onderscheidingsvermogen.

In de wetenschappelijke literatuur is inmiddels veel bekend over het meten van de waarde van informatie (*Value of information*), maar deze kennis wordt weinig toegepast bij het verzamelen van kwaliteitsinformatie in de gezondheidszorg. Uit dergelijk onderzoek kan naar voren komen dat de waarde van sommige informatie zo groot is dat de gebruikelijke betrouwbaarheid van 95% in relatie tot de kosten te laag is. Voor andere informatie worden wellicht(?) onnodig hoge kosten gemaakt.

12. Referenties

- [1] Koolman Advies en Onderzoek (2008), Standaardisatie zorginhoudelijke indicatoren verpleging, verzorging en zorg thuis, (concept 2 mei 2008), Rotterdam
- [2] PwC/TNO (2008) Voorkomen is beter dan genezen. Betrouwbaarheid van kwaliteitsinformatie in de zorg: achtergrondstudie naar risico's en oplossingsrichtingen, Zichtbare Zorg
- [3] Nederlandse Federatie van Universitair Medische Centra's (2008), UMC's gespiegeld.
- [4] Dattalo P, (2008) *Determining Sample Size; Balancing Power, Precision, and Practicality*, Oxford University Press
- [5] Murphy KR, B Myers, A Wolach, *Statistical Power Analysis, A Simple and General Model for Traditional and Modern Hypothesis Tests*, Routledge, New York, derde druk.
- [6] Mant, J. (2001) Process versus outcome indicators in the assessment of quality of health care, *International Journal for Quality in Health Care* 13:475-480
- [7] Gezondheidsraad/Raad voor de Volksgezondheid & Zorg (2006) Vertrouwen in verantwoorde zorg? Effecten van en morele vragen bij het gebruik van prestatie-indicatoren, Centrum voor ethiek en gezondheid.
- [8] Hirshleifer, Jack (1971), The Private and Social Value of Information and the Reward to Inventive Activity, *American Economic Review*, 4: 561-74 pages 1623-1634.

Bijlage I De waarde van statistische betrouwbaarheid

Als gevolg van lage statistische betrouwbaarheid zullen de gebruikers van de prestatie-informatie minder vertrouwen hebben in de informatie. Indien de gebruikers risico-avers zijn, dan zal dit leiden tot een reductie van de waarde van de informatie. Dit pleit voor investeren in betrouwbare informatie. Dit gaat echter veelal gepaard met hoge kosten. Uit figuur 1 blijkt dat een afname van α een groter aantal waarnemingen vereist. De vraag is daarom wat de meerwaarde van extra statistische betrouwbaarheid is (Value of information [8]) en of deze opweegt tegen de meerkosten.

De waarde van betrouwbaarder informatie kan worden geïllustreerd met een voorbeeld. Er zijn drie ziekenhuizen: A, B en C. De voor case-mix gestandaardiseerde kans op sterfte in ziekenhuis A en B is 10% en voor C is 11%. De uitkomsten van ziekenhuis A en C zijn zeer betrouwbaar (precies). De waarneming voor ziekenhuis B is statistisch zeer onbetrouwbaar en kan in de praktijk variëren tussen 0 en 20%. Hoewel ziekenhuizen A en B dezelfde verwachte uitkomst hebben, zullen vele zorgconsumenten de voorkeur geven aan ziekenhuis A boven B omdat zij risico-avers zijn. Een manier om de sterkte van deze voorkeur bij zorgconsumenten te meten, is door zorgconsumenten te vragen of zij C boven B verkiezen.

Het is aannemelijk dat alle gebruikers van kwaliteitsinformatie risico-avers zijn. Hoe risico-avers ze zijn is echter op voorhand niet duidelijk. Juist deze mate van risicoaversie is bepalend voor mate waarin informatie statistisch betrouwbaar moet zijn. Een manier om de waarde van betrouwbare informatie te meten is door te onderzoeken hoeveel geld zorgconsumenten over hebben om de informatie van A te krijgen in plaats van B. Ook de optimale uitruil tussen statistische betrouwbaarheid en onderscheidingsvermogen hangt samen met de voorkeuren van de gebruikers van de informatie.

De waardering van de statistische betrouwbaarheid en het onderscheidingsvermogen van kwaliteitsinformatie kan met behulp van geobserveerd keuzegedrag (revealed preference) en hypothetisch keuzegedrag (stated preference) worden bepaald. De laatste techniek kan weer onderscheiden worden in *contingent valuation* (wat zou u willen betalen indien...) en conjunctanalyse (maak een keuze tussen ziekenhuizen op basis van een aantal eigenschappen en eigen bijdragen). In uitzonderlijke gevallen kan extra onzekerheid waarde opleveren omdat de reële optiewaarde toeneemt met de onzekerheid. Deze toename in waarde is lastig te schatten en vermoedelijk beperkt van omvang.

Bijlage II Berekening van statistische betrouwbaarheid

Het meten van de statistische betrouwbaarheid is mede afhankelijk van de wijze waarop de score op een indicator wordt berekend. Stel dat de statistische betrouwbaarheid moet worden berekend voor de vraag of de geobserveerde score D van organisatie j significant afwijkt van de verwachte score. De geobserveerde score j is de gemiddelde uitkomst van alle waarnemingen n van organisatie j . Voor de ongestandaardiseerde indicator R is de verwachte score gelijk aan de gemiddelde score. Dit gemiddelde is gebaseerd op de som van alle waarnemingen n over j . Dit is gelijk aan N . Formeel:

$$(A.1) \quad R_j = \frac{1}{n} \sum_1^n D_{ij} - \frac{1}{N} \sum_1^N D_{ij}$$

Voor de gestandaardiseerde score S is de verwachte score gelijk aan de gemiddelde verwachte uitkomst op basis van de k case-mix variabelen X voor alle waarnemingen. Formeel:

$$(A.2) \quad S_j = \frac{1}{n} \sum_1^n D_{ij} - \frac{1}{n} \sum_1^n (D_{ij} | X_{ijk})$$

Indien de geobserveerde scores dichotoom zijn (0 of 1) dan wordt de score uitgedrukt in proporties. In dat geval wordt de verwachte uitkomst veelal geschat met behulp van parametrische modellen zoals de logit (logistische) en probit regressiemodellen. Wanneer uitgegaan wordt van een logistisch regressiemodel en de veronderstelling dat S niet is gecorreleerd met X , dan is de significantie van S te benaderen met de significantie van λ behorend bij de indicatorvariabelen Z in de volgende vergelijking:

$$(A.3) \quad \log it(\Pr(D_{ij} = 1 | X_{ijk}, Z_j)) = \alpha_0 + \beta_k X_{ijk} + \lambda_j Z_j$$

Bijlage III Poweranalyse software

Bij het bepalen van de benodigde steekproefgrootte is het gebruikelijk om een poweranalyse te doen. Een poweranalyse stelt de onderzoeker in staat voorafgaand (*a priori*) aan het onderzoek te bepalen of een vooraf bepaald verschil in kwaliteit bij een bepaald aantal observaties significant kan worden aangetoond.

Veel van de informatie die nodig is voor een poweranalyse is vooraf niet helemaal bekend en dient te worden geschat. Het werkelijke statistische betrouwbaarheidsniveau en onderscheidingsvermogen kan daarom achteraf verschillen. Wij raden daarom aan om ook achteraf (*a posteriori* of *post hoc*) een poweranalyse te doen, zodat kan worden vastgesteld of de kwaliteitsinformatie voldoet aan de statistische betrouwbaarheidseisen, of dat de meting dient te worden aangepast.

Er zijn vele softwarepakketen die de gebruiker helpen bij het uitvoeren van een poweranalyse. Zo heeft SPSS een programma voor dit doel ontwikkeld met de naam SamplePower. Dit programma is geschikt voor een t-test voor gemiddelden, tests voor proporties (Chi-square, Fisher's Exact test, McNemar significantie test), correlaties, ANOVA en ANCOVA (tot 3 factoren), regressieanalyse en in het algemeen voor t-, F en Chi-kwadraat verdelingen. Naast SPSS bevatten andere populaire statistische pakketten ook voorgeprogrammeerde routines voor een poweranalyse, waaronder SAS en Stata.

Meer gespecialiseerde programma's zijn G-Power (ook wel G*Power), PASS en 'Power and Precision'. Deze programma's bieden nog enkele aanvullende mogelijkheden. Zo is G-Power in staat om verschillende typen poweranalyse uit te voeren en bevat het programma ook statistische toetsen (onder andere ANOVA met herhaalde metingen en interacties). Veel van de indicatoren zijn dichotoom en worden in de praktijk meestal gestandaardiseerd met behulp van logistische regressie. In die gevallen zijn de software pakketten PASS (NCSS Statistical & Power Analysis Software) en 'Power and Precision' het meest geschikt.

Hoewel de genoemde pakketten speciaal zijn ontwikkeld voor poweranalyse en een groot aantal mogelijkheden hebben, bevatten zij toch vaak niet de specifieke toets die voor prestatie meting wordt gebruikt. Zo is het gebruikelijk om bij een prestatie meting de som van de errortermen per zorgaanbieder uit een regressievergelijking te gebruiken in een berekening. De statistische betrouwbaarheid van dergelijke errortermen kunnen met behulp van simulatietechnieken worden bestudeerd [5]. Voorbeelden van dergelijke technieken zijn Monte Carlo trekkingen en resampling technieken zoals de bootstrap en de jackknife. Deze aanpak vereist een wezenlijke grotere inspanning van de onderzoekers. In dit rapport zijn Monte Carlo trekkingen gebruikt om de verschillen tussen uitkomstschalen te bepalen.

Simulatietechnieken vereisen veel computer en onderzoekstijd en zijn daarom in dit rapport niet vaak gebruikt. De resultaten in dit rapport zijn voornamelijk tot stand gekomen met behulp van Stata (verschillen in proporties en gemiddelden) en met behulp van PASS (logistische regressie).

