# Kleine zorgaanbieders in multilevel vergelijkende analyses

## De CQI Verpleging, Verzorging en Thuiszorg

Dolf de Boer
Lucas van der Hoek
Diana Delnoij
Peter Groenewegen

nivel

ISBN 978-94-6122-017-2

http://www.nivel.nl
nivel@nivel.nl
Telefoon 030 2 729 700
Fax 030 2 729 729

©2010 NIVEL, Postbus 1568, 3500 BN UTRECHT

Niets uit deze uitgave mag worden verveelvoudigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm of op welke andere wijze dan ook zonder voorafgaande schriftelijke toestemming van het NIVEL te Utrecht. Het gebruik van cijfers en/of tekst als toelichting of ondersteuning in artikelen, boeken en scripties is toegestaan, mits de bron duidelijk wordt vermeld.

# Inhoud

# Samenvatting

CQ-index vragenlijsten zijn bedoeld om patiëntervaringen te meten en te vergelijken tussen zorgaanbieders. Voor de betrouwbaarheid van vergelijkingen is het van belang dat er voldoende respondenten zijn per zorgaanbieder, maar wat is voldoende? Vooral in de verpleging verzorging en thuiszorg (VV&T) is dit een belangrijke vraag, want daar is het aantal patiënten per zorgaanbieder vaak beperkt. Het doel van dit rapport was om vast te stellen hoe kleine organisatorische eenheden zich gedragen in multilevel vergelijkende analyses.

Bij vergelijkende analyses voor de CQ-index wordt bekeken of een zorgaanbieder/ organisatorische eenheid zich met zijn gemiddelde en vergelijkingsinterval onderscheid van het populatiegemiddelde. Als dat onderscheid niet wordt aangetoond krijgt de organisatorische eenheid het predikaat ‘gemiddeld’. Het vergelijkingsinterval geeft de onzekerheidsmarge weer rond het gemiddelde van een organisatorische eenheid. Zoals verwacht laat dit rapport zien dat deze onzekerheidsmarge groter is bij organisatorische eenheden met een beperkt aantal respondenten, waardoor onderscheid met het populatiegemiddelde moeilijk aan is te tonen. Daarnaast illustreert dit rapport dat een aantal van de eigenschappen van vergelijkende analyses voor de CQ-index anders uit kunnen pakken voor organisatorische eenheden met weinig respondenten en daarmee ook een onderscheid met het gemiddelde kunnen bemoeilijken.

Wanneer een organisatorische eenheid zich - wegens een beperkt aantal respondenten - niet *kan* onderscheiden van het populatiegemiddelde, dan wordt de classificatie ‘gemiddeld’ voor deze organisatorische eenheid nietszeggend. Kortom, het aantal respondenten voor een organisatorische eenheid moet in ieder geval groot genoeg zijn om een verschil met het populatiegemiddelde aan te kunnen tonen. Of dit het geval is kan worden onderzocht door bij organisatorische eenheden met weinig waarnemingen te bekijken of deze eenheden erin slagen zich te onderscheiden van het gemiddelde. Uit dit rapport blijkt dat het voor eenheden in de VV&T met 10 tot 15 respondenten relatief moeilijk is om zich te onderscheiden, maar in de meeste gevallen is het nog wel mogelijk. Niettemin geldt voor sommige indicatoren dat eenheden van 10 tot 15 respondenten zich niet of nauwelijks kunnen onderscheiden. Dit betekent dat 10 tot 15 respondenten per organisatorische eenheid buitengewoon krap is, maar vanuit pragmatische overwegingen wellicht geaccepteerd kan worden bij eenheden waarbij het niet mogelijk is meer gegevens te verzamelen vanwege hun schaalgrootte. Voor grotere organisatorische eenheden met meer patiënten blijven de steekproefgroottes van 30 respondenten voor interviews, 70 voor vertegenwoordigers en 110 voor de zorg thuis vereist, waarbij dit rapport laat zien dat zelfs de respons die met deze aantallen behaald wordt vanuit wetenschappelijk oogpunt vaak nog als mager wordt beschouwd.

Soms zijn organisatorische eenheden te klein voor vergelijkende analyses, maar behoren zij tot een groter concern van eenheden. In zo'n geval is het de vraag of deze organisatorische eenheden niet bij elkaar gevoegd kunnen worden om zo toch aan het vereiste aantal respondenten te voldoen. Technisch gezien is dit geen enkel probleem, maar conceptueel gezien kleven hier wel gevaren aan. Het is namelijk goed mogelijk dat er aanzienlijke verschillen zijn tussen eenheden van hetzelfde concern, bijvoorbeeld omdat zij beschikken over verschillende faciliteiten of omdat zij zich richten op verschillende patiëntengroepen. Dit betekent dat bij het samenvoegen van eenheden in vergelijkende analyses een hoop informatie verloren kan gaan. Of dit een acceptabel risico is hangt af van de doeleinden waarvoor de resultaten uit vergelijkende analyses worden gebruikt.

De wijze waarop vergelijkende analyses uitpakken voor kleine organisatorische eenheden is deels afhankelijk van de mate waarin eenheden van elkaar verschillen en kan derhalve variëren tussen vragenlijsten. Dit betekent dat enkele van de analyses uit dit rapport opnieuw moet worden uitgevoerd voor andere lijsten alvorens voor die lijsten uitspraken te doen over het minimale acceptabele aantal waarnemingen per organisatorische eenheid.

# 1 Inleiding

De Consumer Quality Index (CQ-index) is een familie van vragenlijsten waarmee de kwaliteit van zorg zoals ervaren door patiënten wordt gemeten. Eén van de doelen van de CQ-index is om de ervaren kwaliteit te vergelijken tussen zorgaanbieders. Een voorbeeld hiervan betreft de landelijke meting in de verpleging, verzorging en thuiszorg (VV&T) waarbij ervaringen zijn gemeten middels interviews of vragenlijsten en vergeleken tussen zorgaanbieders.

Een belangrijke vraag is hoeveel waarnemingen per zorgaanbieder/organisatorische eenheid nodig zijn om tot een betrouwbaar vergelijk te komen. Vooral in de verpleging en verzorging is dit een prangende vraag want daar komt het regelmatig voor dat organisatorische eenheden maar enkele patiënten hebben. Soms zijn dit gewoon kleine zorgaanbieders. In andere gevallen betreft het organisaties van normale grootte die verspreid zijn over verschillende locaties, waardoor zij feitelijk bestaan uit een verzameling kleine organisatorische eenheden. Ten slotte gebeurt het ook wel eens dat een organisatorische eenheid over een redelijk aantal patiënten beschikt, maar dat veel patiënten niet in aanmerking komen voor het onderzoek vanwege de exclusiecriteria en/of dat een aantal vragenlijsten niet in aanmerking komen voor analyse. Dit rapport beoogd inzicht te krijgen in de wijze waarop vergelijkende analyses uitpakken voor kleine organisatorische eenheden met een beperkt aantal waarnemingen. In de volgende alinea's verstrekken we eerst wat achtergrondinformatie over de vergelijkende analyses zoals die voor de VV&T hebben plaats hebben gevonden en komen we vervolgens tot de onderzoeksvragen van dit rapport.

Conform het handboek CQI Meetinstrumenten zijn de vergelijkende analyses voor de VV&T multilevel uitgevoerd. Dit betekent dat er rekening is gehouden met de hiërarchische structuur van de data, d.w.z., respondenten zijn genest in organisatorische eenheden. Wanneer men geen rekening zou houden met deze hiërarchische structuur, dan zouden de statistische analyses gebaseerd zijn op de aanname dat alle respondenten onafhankelijk zijn van elkaar, terwijl dat niet het geval is bij respondenten die tot dezelfde organisatorische eenheid behoren. Verschillen worden dan overschat. Tevens is er een casemix correctie uitgevoerd, maar deze correctie laten we in dit rapport buiten beschouwing omdat we primair geïnteresseerd zijn in hoe kleine organisatorische eenheden zich gedragen in multilevel vergelijkende analyses en niet zozeer in het effect van casemix adjustment daarop. Het effect van casemix adjustment op vergelijkingen tussen organisatorische eenheden staat voor de VV&T elders beschreven (Wiegers et al., 2007; De Boer et al., 2008). Een belangrijk kenmerk van vergelijkende analyses in multilevel modellen betreft de wijze waarop de scores voor organisatorische eenheden tot stand komen. Dit is niet simpelweg het gemiddelde over de waarnemingen van een organisatorische eenheid, maar betreft wat men noemt 'empirical Bayes' (EB) schattingen

(Diez Roux, A. V., 2002). EB schattingen zijn deels afhankelijk van het aantal respondenten binnen een organisatorische eenheid.

## 1.1 Gemiddelden per organisatorische eenheid: empirical Bayes schattingen

In het multilevel model wordt een organisatorische eenheid gezien als onderdeel van een populatie van organisatorische eenheden (Goldstein en Spiegelhalter, 1996). Derhalve zeggen de eigenschappen van de totale populatie organisatorische eenheden dus ook iets over de eigenschappen van een specifieke organisatorische eenheid uit die populatie. Bij het bepalen van gemiddelden per organisatorische eenheid in multilevel modellen wordt dan ook gebruik gemaakt van zowel de informatie van een organisatorische eenheid zelf als informatie over de populatie van organisatorische eenheden als totaal. De informatie van een organisatorische eenheid wordt gecombineerd met informatie over de totale populatie organisatorische eenheden in de vorm van een gewogen gemiddelde. De volgende formule wordt hiervoor gebruikt (Snijders en Bosker, 1999):

$$\beta^{EB}_{0j} = \lambda_j \beta_{0j} + (1 - \lambda_j)\, \gamma_{00} \qquad (1)$$

Waarbij $\beta^{EB}_{0j}$ staat voor het geschatte gemiddelde in organisatorische eenheid $j$ volgens de empirical Bayes methode, $\lambda_j$ staat voor de betrouwbaarheid van het gemiddelde in organisatorische eenheid $j$, $\beta_{0j}$ voor het ruwe gemiddelde in eenheid $j$ en $\gamma_{00}$ voor het gemiddelde over de totale populatie organisatorische eenheden (populatiegemiddelde).

Voor de huidige discussie is de parameter $\lambda_j$ van bijzonder belang, want dit is feitelijk de wegingsfactor die bepaald in hoeverre het populatiegemiddelde wordt vertegenwoordigd in het geschatte gemiddelde van de organisatorische eenheid. De wegingsfactor $\lambda_j$ heeft een range van 0 tot 1 en hoe kleiner $\lambda_j$ hoe meer het populatiegemiddelde is vertegenwoordigd in de schatting van het gemiddelde van een organisatorische eenheid. Met de volgende formule wordt $\lambda_j$ bepaald (Snijders en Bosker, 1999):

$$\lambda_j = \tau^2_0 / (\tau^2_0 + \sigma^2/n_j) \qquad (2)$$

waarbij $\tau^2_0$ staat voor de variantie tussen organisatorische eenheden, $\sigma^2$ staat voor de variantie binnen organisatorische eenheden en $n_j$ het aantal waarnemingen weergeeft binnen organisatorische eenheid $j$. Idealiter is de betrouwbaarheid ($\lambda_j$) groter of gelijk aan 0,80, maar het aantal waarnemingen dat nodig is om een betrouwbaarheid van 0,80 te behalen is niet altijd haalbaar. Uit deze formule volgt ook dat een tweetal zaken invloed heeft op de grootte van $\lambda_j$ en dus op de mate waarin het populatiegemiddelde meeweegt in de schatting van het gemiddelde van een organisatorische eenheid:

Naarmate het aantal waarnemingen van een organisatorische eenheid groter is ($n_j$) wordt $\lambda_j$ ook groter. Het populatiegemiddelde weegt dan minder mee in het geschatte gemiddelde van een organisatorische eenheid.
Wanneer de $n_j$ de waarde '1' aan zou nemen, wordt de formule identiek aan die voor de intraclass correlatie coëfficiënt (ICC; $\tau^2_0 / (\tau^2_0 + \sigma^2)$). Deze coëfficiënt representeert de

verhouding van tussengroepsvariantie en totale variantie. De ICC kan geïnterpreteerd worden als de proportie variantie op het niveau van organisatorische eenheden, maar ook als de verwachte correlatie tussen individuen die tot dezelfde eenheid behoren. Wanneer de ICC groter wordt - er is dan meer variantie op het niveau van de organisatorisch eenheden - zal het populatiegemiddelde minder meewegen in het geschatte gemiddelde van een organisatorische eenheid.

De EB schattingen liggen altijd tussen het populatiegemiddelde en het ruwe gemiddelde (tenzij het ruwe gemiddelde identiek is aan het populatiegemiddelde). EB schattingen worden dan ook wel 'shrinkage estimators' genoemd (Diez Roux, 2002). Bovenstaande determinanten van de wegingsfactor voor het populatiegemiddeld bepalen de *grootte* van deze factor ($\lambda_j$), maar het *effect* van de wegingsfactor wordt mede bepaald door de afstand van het ruwe gemiddelde tot het populatiegemiddelde; hoe kleiner deze afstand hoe minder effect de wegingsfactor heeft. Stel bijvoorbeeld dat het ruwe gemiddelde identiek is aan het populatie gemiddelde, dan maakt het niet meer uit hoe groot de wegingsfactor is, want het gewogen gemiddelde is dan altijd identiek aan zowel het ruwe gemiddelde als het populatiegemiddelde en de weging heeft dan dus geen enkel effect. Kortom, de wegingsfactor ($\lambda_j$) bepaald in hoeverre het populatiegemiddelde vertegenwoordigd is in het geschatte gemiddelde van een organisatorische eenheid en het verschil tussen het ruwe gemiddelde van een organisatorische eenheid en het populatiegemiddelde bepaalt de mate waarin een gegeven wegingsfactor zorgt voor een verschuiving naar het populatiegemiddelde (shrinkage).

De EB schattingen per organisatorische eenheid worden bij CQ-index metingen voorzien van een vergelijkingsinterval (Goldstein en Healy, 1995). Dit vergelijkingsinterval wordt berekend met behulp van de standaard error (de standaard deviatie gedeeld door de wortel uit het aantal waarnemingen) en is dus ook deels afhankelijk van de grootte van een organisatorische eenheid. Deze standaard error wordt vermenigvuldigd met de constante 1,39 en opgeteld en afgetrokken van het EB gemiddelde (Goldstein en Healy, 1995). Vervolgens worden er categorieën gemaakt van relatieve prestaties op basis van deze geschatte gemiddelden met vergelijkingsinterval.

## 1.2 EB schattingen indelen in categorieën: de sterrensystematiek

Om een indicatie te geven van de relatieve prestaties van een organisatorische eenheid worden deze ingedeeld in categorieën van sterren (hoe meer sterren hoe beter). Dit kan een driesterrenindeling zijn, of zoals bij de VV&T, een indeling in vijf sterren. Bij indeling in vijf sterren zijn drie parameters van belang, te weten:

1 het gemiddelde over alle organisatorische eenheden;
2 de gemiddelde bovengrens over alle vergelijkingsintervallen; en
3 de gemiddelde ondergrens over alle vergelijkingsintervallen.

De sterindeling vindt dan als volgt plaats:

***** Het vergelijkingsinterval van de instelling valt geheel boven de gemiddelde bovengrens over alle vergelijkingsintervallen.

**** Het vergelijkingsinterval van de organisatorische eenheid valt geheel boven het gemiddelde over alle organisatorische eenheden, maar niet geheel boven de gemiddelde bovengrens over alle organisatorische eenheden.

*** Het vergelijkingsinterval van de organisatorische eenheid overlapt met het gemiddelde over alle organisatorische eenheden.

** Het vergelijkingsinterval van de organisatorische eenheid valt geheel onder het gemiddelde over alle organisatorische eenheden, maar niet geheel onder de gemiddelde ondergrens over alle organisatorische eenheden.

* Het vergelijkingsinterval van de organisatorische eenheid valt geheel onder de gemiddelde ondergrens over alle vergelijkingsintervallen.

Bij deze sterindeling is het zo dat, hoe groter het vergelijkingsinterval van een organisatorische eenheid, hoe kleiner de kans dat deze eenheid meer of minder dan het gemiddeld aantal sterren krijgt (drie). Organisatorische eenheden die klein zijn hebben over het algemeen vermoedelijk een groter vergelijkingsinterval en zouden zich dus minder makkelijk onderscheiden van het gemiddelde over alle eenheden.

Wanneer kleine organisatorische eenheden met weinig waarnemingen meer opschuiven naar het gemiddelde als gevolg van shrinkage en ook vaker in de driesterren categorie terecht komen, dan heeft dat mogelijk een belangrijk nadeel. Het gevolg is dan namelijk dat wanneer kleine eenheden - als gevolg van schaalgrootte - veel beter of slechter presteren dan grotere eenheden, dit voor een deel wordt weggepoetst door de shrinkage en de sterindeling. Een belangrijke vraag is dan ook of schaalgrootte gerelateerd is aan ervaren kwaliteit.

Ten slotte is het de vraag of EB schattingen in alle gevallen wenselijk zijn. Om dit te kunnen beoordelen is het nuttig om de EB schattingen te vergelijken met ‘gewone’ schattingen zonder shrinkage.

## 1.3 Doel van dit rapport en onderzoeksvragen

Het doel van dit rapport is om in kaart te brengen hoe kleine organisatorische eenheden zich gedragen in vergelijkende analyses. Hiertoe maken we gebruik van de data uit de landelijke meting voor de VV&T. De volgende onderzoeksvragen komen daarbij aan bod:

1 *‘Hoe groot zijn de verschillen tussen de empirical bayes schattingen en de ruwe gemiddelden per organisatorische eenheid?’*

2 *‘Welk aandeel hebben de volgende factoren in het bepalen van het verschil tussen het ruwe gemiddelde en het geschatte ‘empirical Bayes’ gemiddelde: aantal respondenten per organisatorische eenheid, ICC van de betreffende indicator en het verschil tussen het populatiegemiddelde en het ruwe gemiddelde van een organisatorische eenheid?’*

3 *'In hoeverre verschilt de frequentie '3 sterren' in de groep met de kleinste organisatorische eenheden vergeleken met de andere groepen?'*

4 *'Is de grootte van een organisatorische eenheid gerelateerd aan ervaren kwaliteit?'*

5 *'Hoe verhouden de uitkomsten uit de Multilevel analyses met EB schattingen zich tot gewone schattingen zonder shrinkage?'*

# 2 Methode

## 2.1 Vragenlijsten

Voor de cliëntenraadpleging in de VV&T is gebruik gemaakt van de vragenlijst voor interviews met bewoners, de vragenlijst voor vertegenwoordigers van bewoners en de vragenlijst voor thuiswonende patiënten van thuiszorgorganisaties of zorgorganisatorische eenheden. Aangezien de vragenlijsten en de relevante procedures rond dataverzameling en dataverwerking elders uitgebreid zijn beschreven (zie: www.centrumklantervaringzorg.nl) volstaan we hier met een beknopt overzicht van de belangrijkste elementen.

## 2.2 Dataverzameling

De dataverzameling betrof een landelijke meting en werd uitgevoerd door elf commerciële meetbureaus. Vijf van deze bureaus verzamelden data met alledrie de vragenlijsten en nog eens vijf meetbureaus verzamelden uitsluitend data met de vragenlijst voor thuiswonende patiënten van thuiszorgorganisaties of zorgorganisatorische eenheden. Ten slotte was er nog één meetbureau die alleen data verzamelde met de vragenlijst voor vertegenwoordigers van bewoners.

## 2.3 Steekproeftrekking

Bij de interviews met bewoners was het doel om 30 interviews af te nemen. Hiertoe dienden meetbureaus bij de zorgaanbieders een actueel en digitaal bestand op te vragen met daarin alle bewoners, maar niet patiënten van de dagbehandeling, psychogeriatrische patiënten of BOPZ patiënten. Het meetbureau diende hieruit bewoners te verwijderen wanneer één of meer van de volgende exclusiecriteria van toepassing waren: verblijf/ woonduur minder dan één maand, revalidatie of reactivering, kortdurend verblijf (probeerverblijf, respijtzorg of intervalopname), ernstig ziek (zwaarwegende medische redenen), terminale zorg en/of verblijf op palliatieve zorgunit, ernstige psychiatrische problematiek (getraumatiseerd, ernstige gedragsproblemen), dementie (matig-ernstige of ernstige dementie), indicatie psychogeriatrie (PG), andere zwaarwegende redenen (geef een korte omschrijving). Vervolgens moesten meetbureaus twee keer een willekeurige steekproef trekken van 30 patiënten; één voor een selectielijst en één voor een reservelijst. Wanneer bij patiënten van de selectielijst geen data kon worden verzameld, werden patiënten van de reservelijst benaderd. Indien een organisatorische eenheid uit minder dan 30 patiënten bestond, werd de gehele patiëntenpopulatie gemeten, waarbij uiteraard wel dezelfde exclusiecriteria werden gehanteerd als bij grotere organisatorische

eenheden. Uitzonderingen op deze regel vormden eenheden met minder dan 10 patiënten omdat bij dergelijke lage aantallen de anonimiteit in het gedrang komt en dat is ethisch gezien en ivm de privacywetgeving niet te rechtvaardigen. Bij eenheden met minder minder dan 10 bewoners vonden dan ook geen metingen plaats. Van de interviewers werd geëist dat zij getraind en ervaren waren en dat zij de relevante instructies volgden.

Bij de vragenlijst voor vertegenwoordigers van bewoners dienden 70 contactpersonen (vertegenwoordigers) per post te worden benaderd volgens de Dillman methode (Dillman, 2000) welke bestaat uit een eerste verzending en drie herinneringen. Hiertoe moesten meetbureaus een digitaal en actueel bewonersbestand opvragen inclusief contactgegevens van vertegenwoordigers. De afdelingen die met de lijst voor vertegenwoordigers gemeten kunnen worden zijn psychogeriatrische afdelingen, BOPZ afdelingen of algemene afdelingen met een grote groep psychogeriatrische bewoners. Meetbureaus moesten alle bewoners uit het bestand verwijderen als zij geen naaste of familielid als vertegenwoordiger hebben, of als zij slechts kortdurend of minder dan één maand in de zorgorganisatorische eenheid en/of op de betreffende afdeling verblijven. Voorts golden de volgende exclusiecriteria en bewoners die hieraan voldeden werden uit het bestand verwijderd: bewoner heeft geen naaste/familielid, maar alleen een wettelijk vertegenwoordiger; verblijf/woonduur minder dan 1 maand; kortdurend verblijf: probeerverblijf, respijtzorg of intervalopname, terminale zorg en/of verblijf op palliatieve zorgunit; andere zwaarwegende redenen. Uit de overgebleven bewoners diende het meetbureau een willekeurige steekproef van 70 te trekken, tenzij er minder dan 70 bewoners in de database zitten in welk geval voor alle bewoners de relevante contactpersoon werd benaderd waarbij uiteraard wel dezelfde exclusiecriteria werden gehanteerd als bij grotere organisatorische eenheden. Evenals bij de vragenlijst voor interviews met bewoners gold ook hier dat bij eenheden met minder dan tien bewoners geen meting hoefde te worden uitgevoerd vanwege de bescherming van de anonimiteit.

Voor de vragenlijst voor thuiswonende patiënten van thuiszorgorganisaties of zorgorganisatorische eenheden dienden 110 patiënten per post te worden benaderd volgens de Dillman methode (Dillman, 2000). Wederom dienden meetbureaus een actueel en digitaal patiëntenbestand op te vragen. Hieruit moesten zij patiënten selecteren die persoonlijke verzorging en/of verpleegkundige zorg ontvingen, eventueel in combinatie met huishoudelijke verzorging of activerende begeleiding. Patiënten werden geëxcludeerd indien zij jonger waren dan 18 jaar of wanneer zij minder dan zes maanden in zorg waren. Vervolgens moesten meetbureaus een willekeurige steekproef trekken van 110 patiënten tenzij er minder dan 110 patiënten beschikbaar waren in welk geval alle patiënten werden bevraagd die hiervoor in aanmerking kwamen gezien de exclusiecriteria. Wanneer er minder dan 10 patiënten in zorg waren hoefde er niet te worden gemeten vanwege bescherming van anonimiteit.

Meetbureaus dienden ook registratieformulieren in te vullen en te verstrekken aan het CKZ met daarin gegevens over steekproeftrekking, exclusie en respons. Echter de terugkoppeling aan CKZ was op dit punt onvoldoende en een responsanalyse wordt voor dit rapport dan ook niet uitgevoerd. Dit is elders reeds aangemerkt als een verbeterpunt.

## 2.4 Schoning van data

Een uitgebreid overzicht van de opschoningsprocedure is elders beschreven (De Boer et al., 2008). Kort gezegd gaat er bij de interviews met bewoners en vertegenwoordigers van bewoners 3 – 4% van de data verloren omdat er dubbele cases zijn, eenheden die minder dan 10 respondenten aanleverden of omdat één of meerdere exclusiecriteria van toepassing waren blijkens antwoorden op de vragenlijst. Bij de zorg thuis was dit 8,7%. Vervolgens gaat er nog data verloren omdat de variabelen waarvoor gecorrigeerd wordt bij casemix adjustment niet compleet zijn; dit is 5,1% voor de interviews met bewoners, 8,2% voor de vertegenwoordigers van bewoners en 12,7% voor de zorg thuis. Daarnaast kwam het bij de zorg thuis vaak voor dat de vragenlijst was ingevuld en beantwoordt door iemand anders dan de beoogde respondent (13,6% van de gevallen) en ook deze gevallen werden geschoond van de dataset. Na afloop van de schoningsprocedure waren er nog 14.985 respondenten van 577 organisatorische eenheden in de database van interviews met bewoners. Voor de vertegenwoordigers van bewoners waren dit 8.795 respondenten van 306 organisatorische eenheden en voor de thuiszorg waren dit 7.274 patiënten van 221 thuiszorginstellingen.

## 2.5 Berekening van indicatorgemiddelden

Uitgangspunt voor de indicatorberekeningen was het kwaliteitskader verantwoorde zorg (Veen, 2007). Van de vragen die tezamen een indicator vormden is bekeken of zij ook samen een schaal vormden (voor een uitgebreide beschrijving, zie: De Boer et al., 2008). Voor de indicatoren waarbij dat het geval was werd de indicator berekend als het gemiddelde over de onderliggende vragen, maar alleen over respondenten bij wie de helft of minder van de onderliggende vragen missing was. Wanneer de indicator geen schaal vormde, werd één vraag gekozen om de indicator te vertegenwoordigen. Bij de interviews met bewoners gold dit voor indicator 1.1 (ervaringen met zorg [behandel-]/leefplan en evaluatie), indicator 3.2 (ervaringen met maaltijden), indicator 5.1 (ervaren wooncomfort) en indicator 8.1 (ervaren veiligheid woon- leefomgeving). Bij de vertegenwoordigers van bewoners vormden indicator 1.1 (Ervaringen met zorg [behandel-]/leefplan en evaluatie) en indicator 5.1 (ervaren wooncomfort) geen schaal en bij de zorg thuis gold dit voor indicator 1.1 (Ervaringen met zorg [behandel-]/leefplan en evaluatie). Bij de zorg thuis is indicator 5.3 (Ervaren privacy [en woonruimte]) volledig buiten beschouwing gelaten omdat deze voor alle organisatorische eenheden onbetrouwbaar was als gevolg van een coderingsfout door een meetbureau.

## 2.6 Analyses

EB schattingen per organisatorische eenheid zijn gegenereerd met een lineair multilevel regressiemodel waarbij twee niveaus werden onderscheidden, te weten: organisatorische eenheid en respondent. Vergelijkingsintervallen per organisatorische eenheid zijn berekend door de standaarderror te vermenigvuldigen met de constante 1,39 (Goldstein en Healy, 1995) en het resultaat op te tellen bij het gemiddelde van de organisatorische

eenheid voor de bovengrens en af te trekken van het gemiddelde voor de organisatorische eenheid voor de ondergrens.

Om inzichtelijk te maken hoe groot het effect van shrinkage kan zijn is bekeken bij hoeveel procent van de eenheden het ruwe gemiddelde buiten het vergelijkingsinterval van het EB gemiddelde valt. Per vragenlijst is hiervoor een indicator met weinig verschillen tussen eenheden, een indicator met redelijke verschillen tussen eenheden en een indicator met veel verschillen tussen eenheden geselecteerd. Voor de vragenlijst voor interviews en de vragenlijst voor vertegenwoordigers van bewoners waren dit indicator 8.1 (ervaren veiligheid woon- en leefomgeving; weinig verschillen) indicator 2.2 (ervaren informatie; redelijke verschillen) en indicator 5.3 (ervaren privacy; veel verschillen). Voor de zorg thuis waren dit indicator 8.1 (ervaren veiligheid woon- en leefomgeving; weinig verschillen), indicator 4.12 (ervaren professionaliteit en veiligheid zorgverlening; redelijke verschillen) en indicator 9.1 (ervaren beschikbaarheid personeel; veel verschillen). Voor deze selectie van indicatoren is ook onderzocht in hoeverre eenheden met een laag aantal waarnemingen nu vaker in de gemiddelde categorie van drie sterren vallen. Hiertoe zijn eenheden opgedeeld in groepen van verschillende grootte en wordt per groep gerapporteerd hoeveel procent van de eenheden drie sterren kreeg toegewezen.

Om vast te illustreren hoe groot het aandeel van de verschillende determinanten van shrinkage is, zijn lineaire regressieanalyses uitgevoerd met de mate van shrinkage (het verschil tussen het ruwe gemiddelde van een organisatorische eenheid en het EB gemiddelde) als afhankelijke variabele. Als onafhankelijke variabelen zijn het aantal respondenten, de ICC en de afstand tussen het ruwe gemiddelde en het populatiegemiddelde ingevoerd. Voor deze analyses zijn alle indicatoren gebruikt, met uitzondering van indicator 5.3 bij de zorg thuis.

Het aantal waarnemingen per organisatorische eenheid is gecorreleerd aan het ruwe gemiddelde van die eenheid om te bekijken in hoeverre schaalgrootte gerelateerd is aan patiëntervaringen. Hiervoor zijn bivariate Pearson product-moment correlaties uitgevoerd.

Ten slotte zijn schattingen zonder shrinkage berekend door een ANOVA model te schatten met een dummyvariabele voor iedere organisatorische eenheid. Ook voor de ANOVA schattingen zijn vergelijkingsintervallen per organisatorische eenheid berekend door de standaarderror te vermenigvuldigen met de constante 1,39 (Goldstein en Healy, 1995) en het resultaat op te tellen bij het gemiddelde van de organisatorische eenheid voor de bovengrens en af te trekken van het gemiddelde voor de organisatorische eenheid voor de ondergrens.

Emperical Bayes en ICC's schattingen zijn berekend met het softwarepakket MlwiN 2.02. Alle andere analyses zijn uitgevoerd met behulp van SPSS 15.0.

# 3 Resultaten

## 3.1 Het fenomeen ‘shrinkage’

Om meer inzicht te krijgen in de mate van shrinkage is bekeken hoe vaak het ruwe gemiddelde buiten het vergelijkingsinterval van het geschatte gemiddelde valt voor drie indicatoren per lijst. Tabel 3.1 toont de proportie organisatorische eenheden waarbij het ruwe gemiddelde buiten de grenzen van het vergelijkingsinterval van het geschatte gemiddelde valt, per lijst per indicator. Hieruit blijkt dat het percentage organisatorische eenheden waarbij het ruwe gemiddelde buiten de grenzen van het vergelijkingsinterval valt verschilt tussen indicatoren en tussen lijsten. Bij indicatoren met weinig verschillen is de shrinkage het grootst.

Tabel 3.1 Percentage organisatorische eenheden waarbij het ruwe gemiddelde buiten de grenzen van het vergelijkingsinterval valt, uitgesplitst naar vragenlijst en indicator

| | Interviews | | Vertegenwoordigers | | Zorg thuis | |
|---|---|---|---|---|---|---|
| | Indicator | Percentage[a] | Indicator | Percentage[a] | Indicator | Percentage[a] |
| Weinig verschillen | 8.1[b] | 10,7% | 8.1[b] | 12,0% | 8.1[b] | 27,6% |
| Redelijke verschillen | 2.2[c] | 0,9% | 2.2[c] | 6,6% | 4.12[d] | 13,3% |
| Veel verschillen | 5.3[e] | 0,9% | 5.3[e] | 0,0% | 9.1[f] | 5,5% |

[a] percentage organisatorische eenheden waarbij ruwe gemiddelde buiten het vergelijkingsinterval van het geschatte gemiddelde valt
[b] ervaren veiligheid woon- en leefomgeving
[c] ervaren informatie
[d] ervaren professionaliteit en veiligheid zorgverlening
[e] ervaren privacy
[f] ervaren beschikbaarheid personeel

Aangezien shrinkage vrij ingrijpend kan zijn is het van belang vast te stellen waar deze shrinkage door bepaald wordt. Zoals gezegd spelen drie zaken hierin een rol: het aantal waarnemingen waarop een schatting is gebaseerd, de afstand van het ruwe gemiddelde tot het populatiegemiddelde en de intraclass correlatiecoëfficiënt.

## 3.2 Shrinkage: het relatieve aandeel van de ICC, het aantal respondenten per organisatorische eenheid en de afstand van het ruwe gemiddelde van een eenheid tot het populatiegemiddelde

Hoewel de determinanten van shrinkage bekend zijn (het aantal respondenten per organisatorische eenheid, de ICC en de afstand van het ruwe gemiddelde tot het populatiegemiddelde) is het relatieve aandeel van deze determinanten niet eenvoudig te interpreteren op grond van de formules. Ter illustratie zijn daarom lineaire regressieanalyses uitgevoerd met shrinkage als afhankelijke variabele en de ICC, het aantal respondenten per organisatorische eenheid en de afstand van het ruwe gemiddelde van een organisatorische eenheid tot het populatiegemiddelde als onafhankelijke variabelen. Voor deze analyses zijn alle indicatoren gebruikt. Tabel 3.2 toont de resultaten van deze analyses. Wat betreft het aantal respondenten per organisatorische eenheid is te zien dat dit een aanzienlijk effect heeft op de mate van shrinkage: hoe minder respondenten hoe meer de EB schatting opschuift richting het populatiegemiddelde. Ook de ICC is gerelateerd aan de mate van shrinkage en voor twee lijsten zelfs sterker gerelateerd dan het aantal respondenten. Bij een lage ICC schuift de EB schatting meer op naar het populatiegemiddelde. Het verschil tussen het ruwe gemiddelde en het populatiegemiddelde heeft echter verreweg de meeste invloed op de mate van shrinkage. Uitbijters schuiven meer op richting het populatiegemiddelde vergeleken met eenheden die minder extreem scoren.

Tabel 3.2 Lineaire regressies met de mate van 'shrinkage' als afhankelijke variabelen en de determinanten van deze shrinkage als onafhankelijke variabelen (coëfficiënten zijn gestandaardiseerd)

| CQI | ICC | N | Δ (ruwe gemiddelde, populatiegemiddelde) | Degrees of freedom | R2 |
|---|---|---|---|---|---|
| Interviews met bewoners | -0,278* | -0,209* | 0,870* | 3; 8360 | 0,885* |
| Vragenlijst voor vertegenwoordigers | -0,301* | -0,231* | 0,826* | 3; 5505 | 0,796* |
| Vragenlijst voor de zorg thuis | -0,129* | -0,142* | 0,901* | 3; 3093 | 0,875* |

* $p<0,001$

## 3.3 De sterrensystematiek

In tabel 3.3 zijn organisatorische eenheden per indicator uitgesplitst naar het aantal waarnemingen waarop hun geschatte gemiddelde en vergelijkingsinterval zijn gebaseerd voor de interviews met bewoners. Wanneer we kijken naar de kolom 'Frequentie 3 sterren' is te zien dat deze frequentie het hoogste is voor de kleine organisatorische eenheden bij alle drie de indicatoren. Dit komt mogelijk voor een deel door het vergelijkingsinterval, dat ook het grootste is bij de kleinste eenheden. Daarnaast zien we echter ook dat het verschil tussen het geschatte EB gemiddelde en het ruwe gemiddelde ook groter is, oftewel er is meer 'shrinkage'. Aangezien reeds is gebleken dat shrinkage het meest samenhangt met het verschil tussen het ruwe gemiddelde en het gemiddelde

over alle eenheden hebben we dit verschil ook gepresenteerd in tabel 3.2. Dit verschil blijkt over het algemeen inderdaad wat groter bij kleine organisatorische eenheden. Ten slotte is ook de betrouwbaarheid weergegeven in de tabel. Deze betrouwbaarheid (parameter $\lambda_j$ uit formule [2]) wordt gebruikt om de EB schattingen te berekenen en zegt ook iets over de betrouwbaarheid van de schatting. In het algemeen wordt gestreefd naar een betrouwbaarheid van 0,80 of hoger, maar het is niet altijd mogelijk hiervoor voldoende waarnemingen te verzamelen. Voor indicator 2.2 wordt een betrouwbaarheid van 0,80 gehaald als er 25 of meer waarnemingen zijn en voor indicator 5.3 wordt een betrouwbaarheid van 0,80 gehaald vanaf 10 waarnemingen. Een betrouwbaarheid van 0,80 wordt voor indicator 8.1 pas bij benadering gehaald bij 75 waarnemingen (niet in tabel).

Tabel 3.3 Eenheden met een verschillend aantal waarnemingen: frequentie 3 sterren, range vergelijkingsinterval, verschil tussen het ruwe gemiddelde en EB schatting en het verschil tussen het ruwe gemiddelde en het populatie-gemiddelde voor drie indicatoren uit de vragenlijst voor interviews met bewoners

| | Aantal organisatorische eenheden | Frequentie 3 sterren | Breedte vergelijkings-interval | Verschil ruw en EB | Verschil ruw en populatie-gemiddelde | Betrouw-baarheid |
|---|---|---|---|---|---|---|
| **Indicator 8.1[a] (ICC = 4,4%)** | | | | | | |
| 10 t/m 15 waarnemingen | 29 | 100,0% | 0,28 | 0,10 | 0,16 | 0,32 - 0,41 |
| 16 t/m 20 waarnemingen | 64 | 95,3% | 0,26 | 0,07 | 0,13 | 0,41 - 0,48 |
| 21 t/m 25 waarnemingen | 88 | 86,4% | 0,24 | 0,07 | 0,15 | 0,48 - 0,54 |
| 26 t/m 30 waarnemingen | 383 | 80,4% | 0,23 | 0,06 | 0,13 | 0,54 - 0,58 |
| **Indicator 2.2[b] (ICC = 14,3%)** | | | | | | |
| 10 t/m 15 waarnemingen | 67 | 64,2% | 0,55 | 0,11 | 0,37 | 0,63 - 0,71 |
| 16 t/m 20 waarnemingen | 104 | 53,8% | 0,49 | 0,09 | 0,37 | 0,71 - 0,77 |
| 21 t/m 25 waarnemingen | 198 | 49,0% | 0,45 | 0,07 | 0,32 | 0,77 - 0,81 |
| 26 t/m 30 waarnemingen | 182 | 52,2% | 0,42 | 0,06 | 0,29 | 0,81 - 0,83 |
| **Indicator 5.3[c] (ICC = 29,8%)** | | | | | | |
| 11 t/m 15 waarnemingen | 31 | 51,6% | 0,27 | 0,04 | 0,25 | 0,81 - 0,86 |
| 16 t/m 20 waarnemingen | 62 | 24,2% | 0,23 | 0,03 | 0,27 | 0,86 - 0,89 |
| 21 t/m 25 waarnemingen | 97 | 38,1% | 0,21 | 0,02 | 0,22 | 0,89 - 0,91 |
| 26 t/m 30 waarnemingen | 373 | 40,6% | 0,19 | 0,01 | 0,16 | 0,91 - 0,93 |

[a] ervaren veiligheid woon- en leefomgeving
[b] ervaren informatie
[c] ervaren privacy

In tabel 3.4 zien we organisatorische eenheden per indicator uitgesplitst naar het aantal waarnemingen waarop hun geschatte gemiddelde en vergelijkingsinterval zijn gebaseerd voor de vragenlijst voor vertegenwoordigers van bewoners. De resultaten komen sterk

overeen met die uit tabel 3.3: de frequentie driesterren is het hoogste voor de kleine organisatorische eenheden, kleine organisatorische eenheden hebben grotere vergelijkingsintervallen en de mate van shrinkage is ook groter voor kleine organisatorische eenheden. Ook zien we voor indicator 8.1 en indicator 2.2 dat het verschil tussen het ruwe gemiddelde en het populatiegemiddelde groter is bij een klein aantal waarnemingen, maar niet bij indicator 5.3. Bij indicator 8.1 wordt een betrouwbaarheid van 0,80 benaderd bij 50 waarnemingen. Bij indicator 2.2 wordt een betrouwbaarheid van 0,80 gehaald vanaf 40 waarnemingen en bij indicator 5.3 vanaf 10 waarnemingen.

Tabel 3.4 Eenheden met een verschillend aantal waarnemingen: frequentie 3 sterren, range vergelijkingsinterval, verschil tussen het ruwe gemiddelde en EB schatting en het verschil tussen het ruwe gemiddelde en het populatiegemiddelde voor drie indicatoren uit de vragenlijst voor vertegenwoordigers van bewoners

| | Aantal organisato-rische eenheden | Frequentie 3 sterren | Breedte vergelij-kings-interval | Verschil ruw en EB | Verschil ruw en populatie-gemiddelde | Betrouw-baarheid |
|---|---|---|---|---|---|---|
| **Indicator 8.1[a] (ICC = 6,1%)** | | | | | | |
| 10 t/m 20 waarnemingen | 95 | 81,1% | 0,41 | 0,12 | 0,24 | 0,42 - 0,59 |
| 21 t/m 30 waarnemingen | 44 | 70,5% | 0,35 | 0,07 | 0,20 | 0,59 - 0,69 |
| 31 t/m 40 waarnemingen | 47 | 66,0% | 0,31 | 0,05 | 0,19 | 0,69 - 0,74 |
| 41 t/m 50 waarnemingen | 58 | 55,2% | 0,28 | 0,04 | 0,18 | 0,74 - 0,78 |
| **Indicator 2.2[b] (ICC = 9,8%)** | | | | | | |
| 10 t/m 20 waarnemingen | 108 | 67,6% | 0,42 | 0,11 | 0,28 | 0.52 - 0,68 |
| 21 t/m 30 waarnemingen | 43 | 60,5% | 0,35 | 0,06 | 0,23 | 0,68 - 0,76 |
| 31 t/m 40 waarnemingen | 38 | 55,3% | 0,30 | 0,04 | 0,19 | 0,76 - 0,81 |
| 41 t/m 50 waarnemingen | 59 | 50,8% | 0,27 | 0,03 | 0,18 | 0,81 - 0,84 |
| **Indicator 5.3[c] (ICC = 27,2%)** | | | | | | |
| 10 t/m 20 waarnemingen | 108 | 42,6% | 0,47 | 0,05 | 0,30 | 0,79 - 0,88 |
| 21 t/m 30 waarnemingen | 63 | 27,0% | 0,37 | 0,05 | 0,41 | 0,88 - 0,92 |
| 31 t/m 40 waarnemingen | 61 | 16,4% | 0,32 | 0,03 | 0,49 | 0,92 - 0,94 |
| 41 t/m 50 waarnemingen | 36 | 22,2% | 0,29 | 0,02 | 0,29 | 0,94 - 0,94 |

[a] ervaren veiligheid woon- en leefomgeving
[b] ervaren informatie
[c] ervaren privacy

In tabel 3.5 staan de indicatoren voor de zorg thuis weergegeven, wederom uitgesplitst naar het aantal waarnemingen waarop hun geschatte gemiddelde en vergelijkingsinterval zijn gebaseerd. Te zien is dat de kleinste organisatorische eenheden de hoogste frequentie driesterren hebben voor indicator 4.12 en indicator 9.1, maar niet voor indicator 8.1. Verder geldt ook voor de zorg thuis dat de kleinste organisatorische eenheden in alle

gevallen het grootste vergelijkingsinterval hadden en de meeste shrinkage vertoonden. Voor kleinere organisatorische eenheden is ook het verschil tussen het ruwe gemiddelde en het populatiegemiddelde groter. Een betrouwbaarheid van 0,80 wordt voor indicator 4.12 en indicator 9.12 (bij benadering) gehaald bij 50 of meer waarnemingen. Voor indicator 8.1 geldt dat een betrouwbaarheid van 0,80 pas wordt gehaald bij zo'n 125 waarnemingen.

Tabel 3.5 Eenheden met een verschillend aantal waarnemingen: frequentie 3 sterren, range vergelijkingsinterval, verschil tussen het ruwe gemiddelde en EB schatting en het verschil tussen het ruwe gemiddelde en het populatie-gemiddelde voor drie indicatoren uit de vragenlijst voor de zorg thuis

| | Aantal organisato-rische eenheden | Frequentie 3 sterren | Breedte vergelij-kings-interval | Verschil ruw en EB | Verschil ruw en populatie-gemiddelde | Betrouw-baarheid |
|---|---|---|---|---|---|---|
| **Indicator 8.1[a] (ICC = 3,2%)** | | | | | | |
| 10 t/m 20 waarnemingen | 46 | 87,0% | 0,32 | 0,15 | 0,23 | 0,25 - 0,40 |
| 21 t/m 30 waarnemingen | 59 | 93,2% | 0,29 | 0,08 | 0,15 | 0,40 - 0,50 |
| 31 t/m 40 waarnemingen | 46 | 76,1% | 0,27 | 0,08 | 0,17 | 0,50 - 0,57 |
| 41 t/m 50 waarnemingen | 10 | 90,0% | 0,25 | 0,06 | 0,14 | 0,57 - 0,62 |
| **Indicator 4.12[b] (ICC = 6,4%)** | | | | | | |
| 10 t/m 20 waarnemingen | 41 | 78,0% | 0,22 | 0,07 | 0,14 | 0,41 - 0,58 |
| 21 t/m 30 waarnemingen | 33 | 75,8% | 0,19 | 0,05 | 0,12 | 0,58 - 0,67 |
| 31 t/m 40 waarnemingen | 57 | 63,2% | 0,17 | 0,03 | 0,11 | 0,67 - 0,73 |
| 41 t/m 50 waarnemingen | 41 | 75,6% | 0,16 | 0,02 | 0,09 | 0,73 - 0,77 |
| **Indicator 9.1[c] (ICC = 8,0%)** | | | | | | |
| 10 t/m 20 waarnemingen | 43 | 76,7% | 0,28 | 0,07 | 0,17 | 0,47 - 0,64 |
| 21 t/m 30 waarnemingen | 30 | 53,3% | 0,24 | 0,04 | 0,14 | 0,64 - 0,72 |
| 31 t/m 40 waarnemingen | 60 | 46,7% | 0,21 | 0,04 | 0,16 | 0,72 - 0,78 |
| 41 t/m 50 waarnemingen | 37 | 48,6% | 0,19 | 0,03 | 0,14 | 0,78 - 0,81 |

[a] ervaren veiligheid woon- en leefomgeving
[b] ervaren professionaliteit en veiligheid zorgverlening
[c] ervaren beschikbaarheid personeel

Wat betreft shrinkage constateerden we al dat een laag aantal respondenten gepaard gaat met een grotere verschuiving naar het gemiddelde. Dit is onwenselijk wanneer kleine organisatorische eenheden, als gevolg van hun schaalgrootte, consistent hoog of laag scoren. In dat geval is schaalgrootte namelijk gerelateerd aan ervaren kwaliteit en wordt dit deels weggepoetst door de shrinkage estimators. Om hier inzicht in te krijgen hebben we het aantal waarnemingen gecorreleerd aan het ruwe gemiddelde per indicator. Voor de interviews bleek de relatie tussen het aantal waarnemingen en het ruwe gemiddelde te variëren van -0,22 ($p$<0,001; indicator 5.1) tot 0,19 ($p$<0,001; indicator 5.3). Voor de vertegenwoordigers was deze range -0,36 ($p$<0,001; indicator 8.2) tot 0,28 ($p$<0,001;

indicator 5.1) en voor de zorg thuis was dit -0,15 ($p$<0,05; indicator 2.3) tot 0,11 ($p$=0,11; indicator 1.1). Kortom, voor sommige indicatoren geldt dat meer waarnemingen gepaard gaan met hogere ruwe gemiddelden terwijl voor andere indicatoren geldt dat meer waarnemingen gepaard gaan met lagere ruwe gemiddelden. Het is dus mogelijk dat kleine organisatorische eenheden voor sommige indicatoren makkelijker goede zorg kunnen leveren vanwege hun schaalgrootte en voor andere indicatoren minder makkelijk goede zorg kunnen leveren vanwege hun schaalgrootte.

## 3.4 EB schattingen versus schattingen zonder shrinkage

Het is om twee redenen van belang eens te bekijken hoe de multilevel modellen met EB schattingen zich verhouden tot meer gangbare modellen waarbij geen shrinkage wordt toegepast. In de eerste plaats is multilevel analyse een complexe statistische techniek die nog niet in alle software pakketten goed is ingebed. Dit betekent dat het voor onderzoekers die met multilevel modellen willen gaan werken vaak niet eenvoudig is om zich deze modellen eigen te maken. Het is daarom van belang dat de meerwaarde ervan duidelijk wordt gemaakt ten opzichte van eenvoudiger modellen. In de tweede plaats is het wat betreft kleine organisatorische eenheden tot op zekere hoogte de vraag of het toepassen van shrinkage gerechtvaardigd is. Een reden om shrinkage toe te passen is namelijk dat een deel van de verschillen tussen eenheden te wijten kan zijn aan steekproeffouten. En omdat het effect van steekproeffouten groter is bij eenheden met weinig waarnemingen is shrinkage daar ingrijpender. Maar bij kleine eenheden is vaak helemaal geen sprake van een steekproef; daar worden gewoon alle cliënten bevraagd. De vraag is dan ook of shrinkage daar wel moet worden toegepast. Hier komen we in de discussie op terug.

De tabellen 3.6 t/m 3.9 laten zien dat de schattingen uit het ANOVA model een groter betrouwbaarheidsinterval hebben vergeleken met de EB schattingen. Anders gezegd: deze schattingen zijn minder precies. Aangezien de ANOVA schattingen minder precies zijn, zou men verwachten dat er in het ANOVA model ook minder snel significante verschillen worden gevonden tussen organisatorische eenheden. Dit is echter niet het geval: in het ANOVA model zijn er meer eenheden die buiten de drie sterren categorie vallen vergeleken met de EB schattingen uit het multilevel model. Kortom, EB schattingen zijn preciezer én conservatiever dan ANOVA schattingen.

Samengevat laten deze resultaten zien dat eenheden met een klein aantal respondenten zich in Multilevel vergelijkende analyse minder makkelijk kunnen onderscheiden van het populatiegemiddelde om twee redenen:

1 deze eenheden schuiven meer op richting het populatiegemiddelde als gevolg van shrinkage;
2 zij hebben een groter vergelijkingsinterval.

Daarnaast illustreren deze resultaten dat het verschil tussen het ruwe gemiddelde en het populatiegemiddelde veel impact heeft op de mate waarin een organisatorische eenheid opschuift richting het gemiddelde. Kortom, shrinkage is met name ingrijpend voor uitbijters en deze uitbijters lijken vaker voor te komen bij eenheden met een klein aantal

respondenten. Dat shrinkage meer effect heeft op kleine eenheden met een beperkt aantal waarnemingen is onwenselijk als schaalgrootte gerelateerd zou zijn aan ervaren kwaliteit. Deze resultaten tonen dat de relatie tussen het aantal waarnemingen en het ruwe gemiddelde verschilt per indicator. Het is dus voor sommige indicatoren inderdaad niet uitgesloten dat (een gebrek aan) ervaren kwaliteit in kleine eenheden voor een deel wordt weggenomen door shrinkage. Ten slotte is gebleken dat EB schattingen t.o.v. schattingen zonder shrinkage conservatiever zijn (een eenheid scoort eerder gemiddeld) én preciezer zijn (het vergelijkingsinterval van een EB schatting is kleiner).

Tabel 3.6 Interviews met bewoners van verpleeg- of verzorgingshuizen. Emperical Bayes schattingen uit Multilevel analyses versus schattingen uit een ANOVA-model voor eenheden met eenheden met een verschillend aantal waarnemingen: frequentie 3 sterren en range vergelijkingsinterval

| | | Multilevel met shrinkage | | ANOVA | |
|---|---|---|---|---|---|
| | Aantal organisatorische eenheden | Frequentie 3 sterren | Breedte vergelijkings-interval | Frequentie 3 sterren | Breedte vergelijkings-interval |
| **Indicator 8.1[a] (ICC = 4,4%)** | | | | | |
| 10 t/m 15 waarnemingen | 29 | 100,00% | 0,28 | 82,8% | 0,45 |
| 16 t/m 20 waarnemingen | 64 | 95,30% | 0,26 | 78,1% | 0,39 |
| 21 t/m 25 waarnemingen | 88 | 86,40% | 0,24 | 62,5% | 0,34 |
| 26 t/m 30 waarnemingen | 383 | 80,40% | 0,23 | 68,1% | 0,31 |
| **Indicator 2.2[b] (ICC = 14,3%)** | | | | | |
| 10 t/m 15 waarnemingen | 67 | 64,2% | 0,55 | 49,3% | 0,66 |
| 16 t/m 20 waarnemingen | 104 | 53,8% | 0,49 | 47,1% | 0,57 |
| 21 t/m 25 waarnemingen | 198 | 49,0% | 0,45 | 44,9% | 0,50 |
| 26 t/m 30 waarnemingen | 182 | 52,2% | 0,42 | 46,7% | 0,46 |
| **Indicator 5.3[c] (ICC = 29,8%)** | | | | | |
| 11 t/m 15 waarnemingen | 31 | 51,6% | 0,27 | 45,2% | 0,29 |
| 16 t/m 20 waarnemingen | 62 | 24,2% | 0,23 | 21,0% | 0,25 |
| 21 t/m 25 waarnemingen | 97 | 38,1% | 0,21 | 36,1% | 0,22 |
| 26 t/m 30 waarnemingen | 374 | 40,6% | 0,19 | 37,4% | 0,20 |

[a] ervaren veiligheid woon- en leefomgeving
[b] ervaren informatie
[c] ervaren privacy

Tabel 3.7 Vragenlijst voor vertegenwoordigers van bewoners. Emperical Bayes schattingen uit Multilevel analyses versus schattingen uiteen ANOVA-model voor eenheden met eenheden met een verschillend aantal waarnemingen: frequentie 3 sterren en range vergelijkingsinterval

| | | Multilevel met shrinkage | | ANOVA | |
|---|---|---|---|---|---|
| | Aantal organisatorische eenheden | Frequentie 3 sterren | Breedte vergelijkings-interval | Frequentie 3 sterren | Breedte vergelijkings-interval |
| **Indicator 8.1[a] (ICC = 6,1%)** | | | | | |
| 10 t/m 20 waarnemingen | 95 | 81,1% | 0,41 | 69,5% | 0,58 |
| 21 t/m 30 waarnemingen | 44 | 70,5% | 0,35 | 59,1% | 0,43 |
| 31 t/m 40 waarnemingen | 47 | 66,0% | 0,31 | 57,4% | 0,36 |
| 41 t/m 50 waarnemingen | 58 | 55,2% | 0,28 | 51,7% | 0,33 |
| **Indicator 2.2[b] (ICC = 9,8%)** | | | | | |
| 10 t/m 20 waarnemingen | 108 | 67,6% | 0,42 | 54,6% | 0,54 |
| 21 t/m 30 waarnemingen | 43 | 60,5% | 0,35 | 55,8% | 0,41 |
| 31 t/m 40 waarnemingen | 38 | 55,3% | 0,30 | 50,0% | 0,34 |
| 41 t/m 50 waarnemingen | 59 | 50,8% | 0,27 | 47,5% | 0,30 |
| **Indicator 5.3[c] (ICC = 27,2%)** | | | | | |
| 10 t/m 20 waarnemingen | 108 | 42,6% | 0,47 | 38,9% | 0,52 |
| 21 t/m 30 waarnemingen | 63 | 27,0% | 0,37 | 27,0% | 0,39 |
| 31 t/m 40 waarnemingen | 61 | 16,4% | 0,32 | 14,8% | 0,33 |
| 41 t/m 50 waarnemingen | 36 | 22,2% | 0,29 | 22,2% | 0,29 |

[a] ervaren veiligheid woon- en leefomgeving
[b] ervaren informatie
[c] ervaren privacy

Tabel 3.8 Vragenlijst voor de zorg thuis. Emperical Bayes schattingen uit Multilevel analyses versus schattingen uiteen ANOVA-model voor eenheden met eenheden met een verschillend aantal waarnemingen: frequentie 3 sterren en range vergelijkingsinterval

| | | Multilevel met shrinkage | | ANOVA | |
|---|---|---|---|---|---|
| | Aantal organisatorische eenheden | Frequentie 3 sterren | Breedte vergelijkings-interval | Frequentie 3 sterren | Breedte vergelijkings-interval |
| **Indicator 8.1[a] (ICC = 3,2%)** | | | | | |
| 10 t/m 20 waarnemingen | 46 | 87,0% | 0,32 | 60,9% | 0,55 |
| 21 t/m 30 waarnemingen | 59 | 93,2% | 0,29 | 78,0% | 0,43 |
| 31 t/m 40 waarnemingen | 46 | 76,1% | 0,27 | 58,7% | 0,37 |
| 41 t/m 50 waarnemingen | 10 | 90,0% | 0,25 | 50,0% | 0,33 |
| **Indicator 4.12[b] (ICC = 6,4%)** | | | | | |
| 10 t/m 20 waarnemingen | 41 | 78,0% | 0,22 | 58,5% | 0,32 |
| 21 t/m 30 waarnemingen | 33 | 75,8% | 0,19 | 60,6% | 0,24 |
| 31 t/m 40 waarnemingen | 57 | 63,2% | 0,17 | 56,1% | 0,20 |
| 41 t/m 50 waarnemingen | 41 | 75,6% | 0,16 | 68,3% | 0,18 |
| **Indicator 9.1[c] (ICC = 8,0%)** | | | | | |
| 10 t/m 20 waarnemingen | 43 | 76,7% | 0,28 | 72,1% | 0,38 |
| 21 t/m 30 waarnemingen | 30 | 53,3% | 0,24 | 46,7% | 0,29 |
| 31 t/m 40 waarnemingen | 60 | 46,7% | 0,21 | 46,7% | 0,25 |
| 41 t/m 50 waarnemingen | 37 | 48,6% | 0,19 | 45,9% | 0,22 |

[a] ervaren veiligheid woon- en leefomgeving
[b] ervaren informatie
[c] ervaren privacy

# 4 Discussie

Het doel van dit rapport was om vast te stellen hoe kleine organisatorische eenheden zich gedragen in multilevel vergelijkende analyses. Als eerste is bekeken hoe groot de verschillen tussen EB schattingen en ruwe gemiddelden per organisatorische eenheid zijn. Gebleken is dat deze verschillen dusdanig groot uit kunnen vallen dat het ruwe gemiddelde buiten het vergelijkingsinterval van de EB schatting valt. Vervolgens is bekeken waar de mate van shrinkage door werd bepaald. Hoewel de determinanten van shrinkage bekend zijn (het aantal respondenten per organisatorische eenheid, de ICC en de afstand van het ruwe gemiddelde tot het populatiegemiddelde) is het relatieve aandeel van deze determinanten lastig te interpreteren op grond van de formules waarmee shrinkage wordt toegepast. Uit lineaire regressie analyses die ter illustratie zijn uitgevoerd bleek het verschil tussen het ruwe gemiddelde en het populatiegemiddelde verreweg het meest bepalend voor de mate van shrinkage. Dit betekent dat shrinkage vooral veel effect heeft bij instellingen met een extreem lage of een extreem hoge score. Ook de ICC en het aantal respondenten per organisatorische eenheid waren van invloed op de mate van shrinkage, maar minder dan de afstand tussen het ruwe gemiddelde en het populatiegemiddelde.

Voorts is bekeken hoe de indeling in sterren uitpakt voor kleine organisatorische eenheden. Deze sterindeling wijst uit of een organisatorische eenheid zich in positieve of negatieve zin onderscheidt van het gemiddelde. Hiervoor wordt het vergelijkingsinterval van een organisatorische eenheid gebruikt. Wanneer dit interval overlapt met het populatiegemiddelde spreekt men van een gemiddelde score. Bij een vijfsterren indeling krijgt de organisatorische eenheid dan drie sterren toegekend. Men zou verwachten dat het voor kleine organisatorische eenheden met een beperkt aantal waarnemingen moeilijker is om zich te onderscheiden van het gemiddelde omdat zij grotere vergelijkingsintervallen hebben en omdat zij een grotere mate van shrinkage vertonen. Dit bleek inderdaad het geval te zijn: kleine organisatorische eenheden vallen vaker in de categorie drie sterren en dit gaat hand in hand met grotere vergelijkingsintervallen en een grotere mate van shrinkage. Daarnaast viel op dat kleinere organisatorische eenheden met hun ruwe gemiddelde over het algemeen wat verder van het populatiegemiddelde liggen dan grotere organisatorische eenheden. Vermoedelijk komt dit doordat een extreme score meer impact heeft op het gemiddelde in een kleine eenheid met weinig waarnemingen. De mate van shrinkage voor kleine organisatorische eenheden is dus deels groter omdat zich meer uitbijters bevinden onder deze eenheden.

Aangezien shrinkage meer effect heeft op (kleine) eenheden met weinig waarnemingen, is het belangrijk te bekijken of schaalgrootte verband kan houden met ervaren kwaliteit. Uit de correlatie analyse bleek dat verbanden tussen het aantal waarnemingen en het ruwe gemiddelde verschilden per indicator. Voor sommige indicatoren lijken kleine

organisatorische eenheden beter te presteren en voor andere juist slechter. Wat betreft het aantal waarnemingen als maat van de grootte van een organisatorische eenheid moet echter wel een kanttekening worden geplaatst. Het aantal waarnemingen is namelijk niet alleen afhankelijk van de schaalgrootte, maar houdt ook verband met de opschoningsregels en met de respons. Voor eenheden met minder dan 20 waarnemingen kunnen we wel aannemen dat deze eenheden inderdaad klein zijn, maar voor eenheden met meer dan 20 waarnemingen is het minder duidelijk in hoeverre het aantal waarnemingen wordt bepaald door de schaalgrootte of door andere factoren. Hoewel de analyses uit dit rapport dus op hoofdlijnen al een idee geven over de mate waarin schaalgrootte mogelijk gerelateerd is aan ervaren kwaliteit, is het toch ook interessant dit nog eens nader te bestuderen.

Een belangrijke vraag is in hoeverre het wenselijk is dat er shrinkage optreedt in multilevel vergelijkende analyses. Dit rapport heeft laten zien dat als gevolg hiervan verschillen tussen organisatorische eenheden met EB schattingen minder snel naar voren komen vergeleken met andere methoden, een bevinding die overeenstemt met eerder onderzoek (Glance et al., 2006). Enerzijds is dit jammer omdat we juist geïnteresseerd zijn in deze verschillen. Anderzijds betekent dit ook dat de methode conservatiever is dan de meeste andere methoden. Met name bij het schatten van scores voor eenheden met weinig waarnemingen wordt shrinkage beschouwd als een voordeel omdat de beperkte betrouwbaarheid van deze scores wordt ondervangen door mede gebruik te maken van eigenschappen van de totale populatie van eenheden (Arling et al., 2007; Diez Roux, 2002; Zaslavsky, 2001). Echter, bij het toepassen van shrinkage wordt er wel vanuit gegaan dat de waarnemingen waarop de score voor een eenheid is gebaseerd een steekproef vormen van de populatie van die eenheid. Terwijl bij kleine eenheden in de VV&T geen steekproef wordt getrokken, maar de gehele populatie wordt bevraagd. Over de vraag of shrinkage in die gevallen wel zo wenselijk is, kan verschillend worden gedacht. Voorop staat dat de situatie waarin de gehele populatie van een eenheid wordt ondervraagd niet hetzelfde is als de situatie waarin een steekproef wordt bevraagd. Anderzijds kan men redeneren dat de populatie van een kleine eenheid op ieder moment kan veranderen. De cliënten in een kleine eenheid vormen dus een steekproef van de cliënten die er hadden kunnen zitten. Een vergelijkbare redenering betreft de stelling dat de ervaringen die in een kleine eenheid worden gemeten een steekproef vormen van alle mogelijke ervaringen die de betreffende cliënten hadden kunnen hebben of rapporteren. Kortom, hoewel het bevragen van de gehele populatie anders is dan het bevragen van een steekproef, zijn er in beide gevallen goede argumenten voor het toepassen van shrinkage.

Het is in het algemeen lastig een duidelijke grens te identificeren voor een minimaal aantal waarnemingen bij vergelijkende analyses. Vanuit statistisch oogpunt zou het wenselijk zijn voor alle indicatoren een betrouwbaarheid van ten minste 0,80 te hebben. Dit betekent dat er 50 complete interviews met bewoners beschikbaar moeten zijn per verpleeg- of verzorgingshuis. Bij de huidige eis van 30 complete interviews per verpleeg- of verzorgingshuis wordt de gewenste betrouwbaarheid van 0,80 dus niet voor alle indicatoren gehaald. Voor de vertegenwoordigers van bewoners geldt dat bij een verwachte respons van 78% (Wiegers et al., 2007) er 64 vertegenwoordigers moeten worden aangeschreven om de vereiste 50 waarnemingen per verpleeg- of verzorgingshuis

te behalen. De huidige richtlijn waarbij 70 vertegenwoordigers worden aangeschreven is dus voldoende. Bij een verwachte respons van 52% voor de thuiszorg (Wiegers et al., 2007) geldt dat er 240 patiënten moeten worden aangeschreven om de vereiste 125 waarnemingen per thuiszorgorganisatie te behalen. De huidige richtlijn waarbij 110 patiënten worden aangeschreven is dus niet voldoende om een betrouwbaarheid van 0,80 te halen voor alle indicatoren in de thuiszorg. In de praktijk wordt vaak pragmatisch omgegaan met de wens dat de betrouwbaarheid 0,80 of hoger moet zijn. Doorgaans is men bereid de steekproefgrootte wat naar beneden bij te stellen en te accepteren dat de betrouwbaarheid niet voor alle indicatoren 0,80 of hoger is. Daarnaast komt het regelmatig voor dat er bij de kleine(re) organisatorische eenheden niet voldoende patiënten beschikbaar zijn om de gewenste steekproefgroottes te behalen, waardoor de betrouwbaarheid soms fors lager ligt dan de gewenste 0,80. Men zou kunnen besluiten bij deze eenheden geen meting uit te voeren, maar dat gebeurt vaak niet omdat het beleidsmatig uitgangspunt is dat ook deze eenheden zich dienen te verantwoorden. Het is dan ook van belang in kaart te brengen hoe de vergelijkende analyses uitpakken voor deze kleine eenheden.

Dit rapport heeft zich vooral gericht op de huidige praktijk in de VV&T waarbij vergelijkende analyses plaatsvinden wanneer een organisatorische eenheid ten minste 10 respondenten aanlevert. Bekeken is of kleine eenheden zich, gezien de grotere vergelijkingsintervallen en grotere mate van shrinkage nog kunnen onderscheiden van het populatiegemiddelde. Voor deze eenheden bleek het inderdaad lastig zich te onderscheiden van het gemiddelde, maar in de meeste gevallen was het niet onmogelijk. Alleen bij indicatoren waarbij in het algemeen weinig verschillen tussen eenheden naar voren komen was het voor eenheden met 10 tot 15 respondenten niet altijd mogelijk zich te onderscheiden. Een gemiddelde score van drie sterren is voor eenheden met 10 tot 15 waarnemingen dan ook minder informatief omdat het voor deze eenheden relatief moeilijk is om zich te onderscheiden buiten deze categorie te vallen. Maar wat betekent dit nu allemaal voor het minimum aantal waarnemingen voor vergelijkende analyses in de VV&T? Het lijkt er in ieder geval op dat dit minimum niet onder de 10 gezocht moet worden. Dit is om te beginnen al niet mogelijk omdat met zo weinig waarnemingen de anonimiteit in het gedrang komt, maar daarnaast is het voor eenheden met 10 tot 15 waarnemingen al dusdanig lastig om zich te onderscheiden van het gemiddelde dat de resultaten van deze eenheden al beperkt informatief zijn. Niettemin is het minimum van 10 waarnemingen per organisatorische eenheid een interessante optie omdat hiermee niet al te veel eenheden afvallen en omdat het voor eenheden met 10 of meer waarnemingen over het algemeen nog wel mogelijk is zich te onderscheiden. Het is hierbij van het grootste belang te beseffen dat een minimum van 10 waarnemingen vanuit wetenschappelijk oogpunt veel te krap is. Derhalve bevelen we aan om resultaten van eenheden die dicht bij het minimum aantal waarnemingen zitten extra kritisch te bekijken, omdat de betrouwbaarheid in deze gevallen veelal lager is dan vanuit wetenschappelijk oogpunt is vereist.

Soms zijn organisatorische eenheden te klein voor vergelijkende analyses, maar behoren zij tot een grotere groep van eenheden. In zo'n geval is het de vraag of deze organisatorische eenheden niet bij elkaar gevoegd kunnen worden om zo toch aan het

vereiste aantal respondenten te voldoen. Technisch gezien is dit geen enkel probleem, maar conceptueel gezien kleven hier wel gevaren aan. Het is namelijk goed mogelijk dat er aanzienlijke verschillen zijn tussen eenheden van hetzelfde concern, bijvoorbeeld omdat zij beschikken over verschillende faciliteiten of omdat zij zich richten op verschillende patiëntengroepen. De verschillen in patiëntervaringen die hiermee gepaard gaan kunnen voor een groot deel verdwijnen bij een samenvoeging van eenheden van hetzelfde concern. Anders gezegd: bij het samenvoegen van eenheden in vergelijkende analyses kan een hoop informatie verloren gaan(Bird et al., 2005). Of dit een acceptabel risico is hangt af van de doeleinden waarvoor de resultaten uit vergelijkende analyses worden gebruikt. Een oplossing kan zijn een dergelijke samenvoeging wel toe te staan en deze te voorzien van een kanttekening die door de verschillende stakeholders naar eigen inzicht gewogen kan worden. Niettemin zal het altijd lastig blijven om data over samengevoegde eenheden te interpreteren.

Men zou zich af kunnen vragen of het minimaal acceptabele aantal respondenten voor vergelijkende analyses niet gewoon vooraf bepaald had kunnen worden met een poweranalyse. Uiteraard is het zinnig om vooraf poweranalyses te doen. Uitgangspunt hierbij is doorgaans het minimale verschil dat men wil kunnen vinden (bijvoorbeeld een verschil van 0,5 in het rapportcijfer). Vervolgens wordt dan een formule toegepast waarmee het aantal benodigde respondenten per organisatorische eenheid wordt berekend om een verschil van 0,5 in rapportcijfer in 80% van de gevallen te kunnen aantonen. Door het aantal benodigde respondenten te combineren met de verwachte respons wordt bepaald hoeveel patiënten moeten worden benaderd (wanneer 10 respondenten benodigd zijn, dienen bij een verwachte respons van 50% 20 patiënten te worden benaderd). Echter, deze strategie is erop gericht het aantal patiënten dat benaderd wordt aan te passen aan het minimale verschil dat men wil kunnen vinden en dit kan makkelijk leiden tot aantallen die groter zijn dan haalbaar is voor kleine eenheden. Inschattingen op basis van poweranalyses zijn daarnaast deels ook gebaseerd op de verschillen die men verwacht te vinden. Wanneer achteraf blijkt dat de verschillen veel kleiner of juist veel groter zijn dan verwacht zal toch opnieuw moeten worden bekeken of het voor eenheden met een beperkt aantal waarnemingen nog mogelijk is om zich van anderen te onderscheiden. Ten slotte wordt het fenomeen shrinkage niet meegenomen in de poweranalyses zoals die nu voor de CQ-index worden uitgevoerd (Sixma en Delnoij, 2007). Hoewel het dus nuttig is om vooraf een poweranalyse te doen voor een globale indruk van het aantal mensen dat moet worden aangeschreven, blijft het achteraf ook de moeite waard om te bekijken hoe de vergelijkende analyses en de sterindeling uitpakken voor kleine eenheden met weinig waarnemingen.

# 5 Conclusies

Kleine organisatorische eenheden hebben in vergelijkende analyses moeite zich te onderscheiden van het gemiddelde. Dit komt omdat zij doorgaans grotere vergelijkingsintervallen hebben en omdat zij gevoeliger zijn voor ‘shrinkage’. Voor eenheden met 10 tot 15 waarnemingen is het over het algemeen niet makkelijk, maar wel mogelijk zich te onderscheiden van het gemiddelde. Een minimum van 10 waarnemingen is daarom wellicht acceptabel voor eenheden waarbij niet meer data verzameld kan worden maar meer is duidelijk vereist in de gevallen waar dat mogelijk is. Het is hierbij belangrijk op te merken dat het aanleveren van 10 respondenten aan de centrale database geen garantie is voor de aanwezigheid van 10 waarnemingen voor iedere indicator. Het gebeurt namelijk regelmatig dat meetbureaus data aanleveren waarvan nog respondenten verloren gaan bij het toepassen van de schoningsregels. Daarnaast komt het vaak voor dat een respondent niet meeweegt voor een indicator omdat teveel van de vragen waaruit deze indicator bestaat waren beantwoordt met ‘niet van toepassing’ of ‘weet ik niet’.

Voor toekomstige metingen en ook voor andere lijsten is het van belang enkele van de analyses uit dit rapport te herhalen alvorens een besluit te nemen over het minimale aantal waarnemingen dat geaccepteerd kan worden voor eenheden die een kleine patiënten-populatie hebben. De meest eenvoudige strategie is om voor een groep eenheden met een beperkt aantal respondenten te bekijken of zij erin slagen zich binnen de sterindeling te onttrekken van de gemiddelde categorie. Hierbij dient wel duidelijk te zijn dat zo’n analyse gericht is op het minimaal acceptabele aantal waarnemingen *in gevallen waarin niet kan worden voldaan aan de vereiste steekproefgrootte*. Wanneer dit minimum (van bijvoorbeeld 10 respondenten) zou worden doorgevoerd in de steekproeftrekking voor alle eenheden, dan is dat zeer bezwaarlijk om twee redenen. Ten eerste levert dit voor eenheden waarbij meer patiënten beschikbaar zijn gegevens op die veel minder informatief zijn dan mogelijk is. Ten tweede is het voor de analysemethoden van belang dat er in zijn totaliteit veel eenheden en respondenten beschikbaar zijn, want anders worden de gebruikte modellen minder stabiel en zijn de uitkomsten minder betrouwbaar. Het minimaal acceptabele aantal waarnemingen in gevallen waarin niet kan worden voldaan aan de vereiste steekproefgrootte mag dus nooit een argument zijn om de vereiste steekproefgrootte ook voor grotere eenheden te verlagen.

# Literatuur


Arling G, Lewis T, Kane RL, Mueller C, Flood S. Improving quality assessment through multilevel modeling: the case of nursing home compare. *Health Serv Res,* 2007; 42:1177-99

Bird S, Cox D, Farewell V, Goldstein H, Holt T, Smith P. Performance indicators: good, bad and ugly. *J R Stat Soc Ser A Stat Soc*, 2005; 168:1-27

Boer D de, Damman OC, Delnoij D. *Meetverantwoording cliëntgebonden indicatoren VV&T.* Utrecht: Centrum Klantervaring Zorg, 2008

Diez Roux AV. A glossary for multilevel analysis. *J Epidemiol Comm Health,* 2002; 56:588-94

Dillman DA. *Mail and internet surveys: the Tailored Design Method.* New York: John Wiley Co., 2000

Glance LG, Dick A, Osler TM, Li Y, Mukamel DB. Impact of changing the statistical methodology on hospital and surgeon ranking: the case of the New York State cardiac surgery report card. *Med.Care,* 2006; 44:311-9

Goldstein H, Healy MJR. (1995). The Graphical Presentation of a Collection of Means. *J R Stat Soc Ser A Stat Soc*, 1995; 158:175-7

Goldstein H, Spiegelhalter DJ. League Tables and their Limitations: Statistical Issues in Comparisons of Institutional Performance. *J R Stat Soc Ser A Stat Soc*, 1996; 159:385-443

Sixma H, Delnoij D. *Handboek CQI meetinstrumenten: een handleiding voor de ontwikkeling en het gebruik van Consumer Quality Index (CQI) vragenlijsten.* Utrecht: NIVEL, 2007

Snijders TAB, Bosker RJ. *Multilevel Analysis: An introduction to basic and advanced multilevel modeling.* Londen: Sage Publishers, 1999

Veen JAH. Kwaliteitskader verantwoorde zorg. Utrecht: Stuurgroep Verantwoorde Zorg VV&T, 2007

Wiegers TA, Stubbe JH, Triemstra M. *Ontwikkeling van een CQ-Index voor verpleeg- en verzorgingshuizen en thuiszorg: kwaliteit van zorg volgens bewoners.* Utrecht: NIVEL, 2007

Zaslavsky AM. Statistical issues in reporting quality data: small samples and casemix variation. *Int J Qual Health Care,* 2001; 13:481-8