



Dit rapport is een uitgave van het NIVEL.  
De gegevens mogen worden gebruikt met  
bronvermelding.

**Versnelling van vergelijkende analyse en casemixcorrectie**  
Een geautomatiseerde procedure met behulp van een referentiedataset

Dolf de Boer  
Lucas van der Hoek

U vindt dit rapport en andere publicaties van het NIVEL in PDF-format op: [www.nivel.nl](http://www.nivel.nl)

ISBN 978-94-6122-285-5

<http://www.nivel.nl>

[nivel@nivel.nl](mailto:nivel@nivel.nl)

Telefoon 030 2 729 700

Fax 030 2 729 729

©2014 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

<b>1 Inleiding</b>	<b>7</b>
1.1 Achtergrond	7
1.2 Werkwijze bij grootschalige metingen	8
1.3 Een alternatieve en geautomatiseerde procedure	8
1.4 Doel van het rapport	9
<b>2 Methode</b>	<b>11</b>
2.1 Vragenlijsten en indicatoren	11
2.2 Dataverwerking	11
2.2.1 Aanlevering, schoning en referentiedatasets	11
2.2.2 Analyse, casemixcorrectie en sterrenberekening	12
2.2.3 Representativiteit van referentiedatasets	14
<b>3 Resultaten</b>	<b>17</b>
3.1 Kenmerken referentiedata versus overige data	17
3.2 Indicatorscores referentiedata versus overige data	19
3.3 Invloed van de referentiedataset op de sterrenclassificatie	20
3.4 Kleinere referentiedatasets	20
<b>4 Discussie</b>	<b>23</b>
4.1 Kernbevindingen	23
4.2 Ervaringen met het geautomatiseerde proces: aandachtspunten	24
4.3 Referentiedatasets: eisen en onderhoud	24
4.4 Conclusie	25
<b>Literatuur</b>	<b>27</b>



## Voorwoord

De resultaten van kwaliteitsmetingen zijn beter bruikbaar naarmate deze actueler zijn. Resultaten uit grootschalige landelijke metingen komen om verschillende redenen echter vaak pas vele maanden na de metingen beschikbaar. Eén van die redenen betreft de toepassing van state-of-the-art methoden voor vergelijkende analyses en casemixcorrectie.

In dit rapport staat een methode beschreven waarbij, met behulp van een referentiedataset, vergelijkende analyses met casemixcorrectie veel sneller uitgevoerd kunnen worden dan in het verleden. De methode is geautomatiseerd en ontwikkeld voor de Consumer Quality Index Verpleging, Verzorging en Zorg thuis (CQ-index; VVT). De CQ-Index is een familie van vragenlijsten voor het meten van patiëntervaringen met bijbehorende methodiek. De methode die in dit rapport wordt geïntroduceerd voldoet aan de eisen die horen bij de CQ-Index. De methode kan worden toegepast in sectoren waarin op basis van lopende of vorige metingen een referentiedataset van enkele tientallen aanbieders kan worden geconstrueerd.

Naast de mogelijkheid om analyses te versnellen heeft de methode met gebruik van referentiedatasets nog enkele andere voordelen die in dit rapport beschreven staan. Daarnaast verkent dit rapport de eventuele eisen die aan referentiedatasets gesteld kunnen worden, waarbij onder meer wordt gekeken naar de omvang en de representativiteit van de referentiedataset.

De methode draait inmiddels meer dan een half jaar in de VVT en wordt gebruikt om casemixgecorrigeerde scores toe te voegen aan de spiegelrapportage die brancheorganisatie ActiZ beschikbaar stelt aan zorgaanbieders en cliëntenraden. Indicatoren die worden geanalyseerd op dezelfde of vergelijkbare wijze als bij de CQ-index kunnen veel baat hebben bij de hier beschreven methode voor het versnellen van dit soort analyses. Dit verhoogt de actualiteit en de bruikbaarheid van de resultaten.

NIVEL, december 2014



# 1 Inleiding

## 1.1 Achtergrond

Transparantie van kwaliteit is van groot belang voor het borgen of verbeteren van kwaliteit van zorg (Ministerie VWS, 2013). Transparantie van kwaliteit helpt zorgaanbieders mogelijkheden voor interne kwaliteitsverbetering te identificeren en kan ook cliënten/patiënten helpen bij de keuze voor een zorgaanbieder. Tevens kunnen verzekeraars informatie over kwaliteit van zorg gebruiken bij de zorginkoop. In dit licht zijn de afgelopen jaren vele indicatoren ontwikkeld die inzicht moeten bieden in kwaliteit van zorg. Een deel van die indicatoren is gebaseerd op de Consumer Quality Index (CQ-index of CQI), een methode voor het meten van cliëntervaringen met kwaliteit van zorg.

Doorgaans wordt de kwaliteit van zorg inzichtelijk gemaakt door de scores van zorgaanbieders op indicatoren onderling te vergelijken, het zogeheten ‘benchmarken’. De resultaten worden gebruikt voor interne kwaliteitsverbetering, maar komen ook beschikbaar in het publieke domein voor externe verantwoording, cliëntkeuze en zorginkoop. Met name de publieke toegankelijkheid van scores op kwaliteitsindicatoren stelt hoge eisen aan de analyses die worden gebruikt om deze scores te berekenen. Voor indicatoren gebaseerd op de CQ-index wordt hierbij gebruik gemaakt van multilevel regressie modellen (zie box 1.1) waarin wordt gecorrigeerd voor casemix (Sixma et al., 2008; CKZ, 2011; De Boer et al., 2010). Deze benadering wordt voor andere indicatoren in Nederland ook steeds meer toegepast (Koolman et al., 2012) en is internationaal ook geaccepteerd als goede methode voor het vergelijken van zorgaanbieders (Arling et al., 2007; Damman et al., 2009; Goldstein en Spiegelhalter, 1996).

### Box 1.1 Overwegingen bij de keuze voor multilevel regressie

Er zijn twee belangrijke redenen voor de keuze voor multilevel regressie als analysetechniek voor het vergelijken van zorgaanbieders:

- 1 Een gewone (single-level) regressie gaat er van uit dat ieder individu evenveel unieke informatie toevoegt. Bij het vergelijken van zorgaanbieders is dat echter niet reëel omdat individuen van dezelfde zorgaanbieder meer gemeenschappelijk hebben dan individuen van verschillende zorgaanbieders. In een multilevel regressie wordt hier rekening mee gehouden (Snijders & Bosker, 1999).
- 2 Het effect van toevalsvariatie bij zorgaanbieders met weinig waarnemingen wordt geminimaliseerd m.b.v. empirical Bayes schattingen (Greenland, 2000; Arling et al., 2007; de Boer et al., 2010).

## 1.2 Werkwijze bij grootschalige metingen

Voor zorgaanbieders is het van groot belang zo snel mogelijk te beschikken over de resultaten zodat deze direct kunnen worden ingezet voor kwaliteitsverbetering. Ook andere gebruiksdoelen zijn gediend met een snelle terugkoppeling van metingen van kwaliteit van zorg. Vanuit methodisch oogpunt is echter de koninklijke weg om alle data van alle instellingen in één keer te analyseren. Dit zorgt er namelijk voor dat alle parameters die een rol spelen in vergelijkende (benchmark-) analyses en casemixcorrectie (regressiecoëfficiënten, varianties, et cetera) in één keer worden vastgesteld op grond van een zo compleet mogelijke dataset. De analyses kunnen dan pas plaatsvinden als alle data binnen zijn, waarmee de aanbieders die er vroeg bij zijn soms vele maanden moeten wachten op aanbieders waarbij de metingen trager verlopen. In de Verpleging, Verzorging en Zorg thuis is er om die reden voor gekozen de analyses in batches van bijvoorbeeld eens per jaar uit te voeren op grond van de data die in die periode zijn aangeleverd.

Ondanks de verwerking in batches bleef de doorlooptijd van metingen, aanlevering van data, controles, schoning en analyse, accordering van resultaten door zorgaanbieders en openbaarmaking een knelpunt (Hopman et al., 2011). Daarnaast was de constructie vanuit methodisch oogpunt niet geheel bevredigend omdat verschillende parameters die een rol spelen in de analyse bij iedere batch opnieuw werden vastgesteld. Voor grensgevallen – aanbieders die bijvoorbeeld *nét* wel of *nét* niet bovengemiddeld scoren – kan het dan uitmaken binnen welke batch zij worden geanalyseerd. Dat arbitraire keuzes in het dataverwerkingsproces voor grensgevallen net het verschil kunnen maken tussen een gemiddelde vs een beneden- of bovengemiddelde score is overigens nooit helemaal te vermijden, maar dient natuurlijk wel zoveel mogelijk te worden voorkomen.

## 1.3 Een alternatieve en geautomatiseerde procedure

Een alternatief voor het analyseren in batches is om te werken met een vaste referentiedataset waar nieuwe gegevens van een aanbieder aan worden toegevoegd. Vervolgens vinden de analyses en casemixcorrectie plaats waarna resultaten van die ene aanbieder worden doorgeleverd en de data van die aanbieder weer uit de referentiedataset worden gehaald. Dit alternatief lost twee problemen op: (1) Zodra een referentiedataset beschikbaar is kan nieuwe data op ieder gewenst moment worden geanalyseerd (in plaats van te wachten tot een nieuwe batch groot genoeg is voor een deugdelijke analyse) en (2) zolang de referentiedataset constant blijft is er geen sprake meer van dat een aanbieder op het ene moment of in de ene batch net wel boven- of ondergemiddeld scoort en in de andere batch of op het andere moment net niet. Deze werkwijze is in beginsel echter wel buitengewoon bewerkelijk. Het vereist namelijk een aparte analyse voor iedere aanbieder die data aanlevert. Om dit hanteerbaar te maken is het dan ook wenselijk, zo niet noodzakelijk, om dit te automatiseren.

Een belangrijke stap bij het gebruik van een vaste referentiedataset is het samenstellen van deze referentieset. Een pragmatische werkwijze is om de eerste groep zorgaanbieders



die aanleveren als referentiegroep te nemen. Dan kan de procedure worden ingericht zodra de referentiegroep groot genoeg is. De vraag is wel of zo'n eerste groep zorgaanbieders representatief is, en zo nee, wat de consequenties zijn als de referentiegroep niet (geheel) representatief is voor de totale populatie.

#### **1.4 Doel van het rapport**

In dit rapport wordt verslag gedaan van vergelijkende analyses en casemixcorrectie voor de CQ-index Verpleging, Verzorging en Zorg Thuis (VVT) met het gebruik van een referentiedataset. De procedure is geautomatiseerd en de ontwikkeling van de procedure en automatisering zijn uitgevoerd in samenwerking met ActiZ ten behoeve van de spiegelrapportage die ActiZ gratis aanbiedt aan haar leden en aan cliëntenraden en waarin de CQ-index een belangrijke rol speelt. De referentiedataset is gebaseerd op de eerste groep zorgaanbieders die heeft aangeleverd. De doelen van dit rapport zijn als volgt:

- 1 Het beschrijven van (de ervaringen met) de versnelde en geautomatiseerde procedure van vergelijkende analyses en casemixcorrectie met behulp van een referentiedataset, teneinde de opgedane kennis zoveel mogelijk te verspreiden.
- 2 Vaststellen in welke mate de groep zorgaanbieders uit de referentiedataset verschilt van de totale dataset en in welke mate dit van invloed is op de resultaten van een individuele aanbieder.

In hoeverre het al dan niet een probleem is als de groep zorgaanbieders uit de referentiedataset niet (geheel) representatief is voor alle zorgaanbieders komt in de discussie aan de orde.



## 2 Methode

### 2.1 Vragenlijsten en indicatoren

De CQ-index VVT bestaat uit drie vragenlijsten: een vragenlijst voor interviews met cliënten van verpleeg- en verzorgingshuizen (CQI VV), een vragenlijst voor vertegenwoordigers van cliënten met psychogeriatrische problematiek (CQI PG) en een vragenlijst voor cliënten die zorg thuis ontvangen (CQI ZT). De data uit dit rapport zijn gebaseerd op versie 4.1 van de vragenlijsten (CKZ, 2012).

De dataverzameling met deze vragenlijsten is uitgevoerd door geaccrediteerde meetbureaus. De richtlijnen en werkinstructies van het voormalige Centrum Klantervaring Zorg waren van toepassing op de dataverzameling (CKZ, 2012). De privacy van respondenten is binnen deze werkinstructies geborgd. Bij het berekenen van indicatoren zijn eveneens richtlijnen en werkinstructies van het CKZ gevolgd die ook aansluiten daarmee aan bij het kwaliteitsdocument VVT (IGZ, ZN en LOC, 2013).

### 2.2 Dataverwerking

#### 2.2.1 Aanlevering, schoning en referentiedatasets

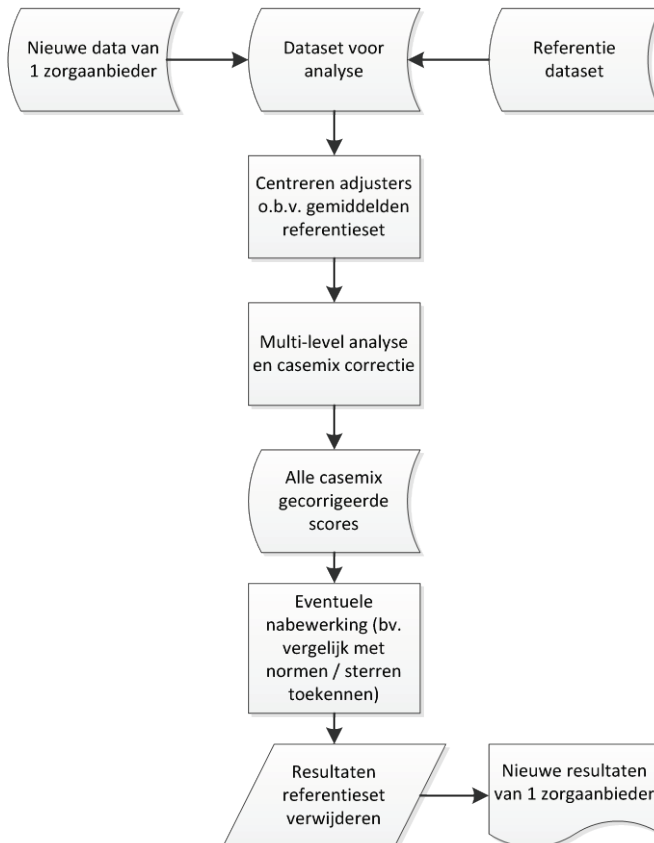
Op verzoek van de zorgaanbieders leveren meetbureaus gegevens aan bij één partij, die namens ActiZ een portal beheert ten behoeve van de spiegelrapportage. Het schonen van data vond plaats bij aanlevering aan deze portal, waarbij de schoningsregels voor de CQI VVT alsmede verschillende andere controles werden ingebouwd. De partij leverde vervolgens de data aan bij het NIVEL voor de casemix correctie. Vanuit het vier-ogen principe heeft het NIVEL een eenmalige controle van de schoning uitgevoerd op de eerste grote datalevering die door de partij werd geleverd. Het NIVEL berekende de indicatoren zoals omschreven in het Kwaliteitsdocument VVT (IGZ, ZN en LOC, 2013). Wanneer deze uit meerdere vragen bestaan zijn alleen de cliënten meegenomen die de helft of meer van de vragen hebben ingevuld. Twee indicatoren bestonden uit één vraag met twee antwoordcategorieën en vereisten een logistisch model. Deze zijn buiten beschouwing gelaten voor dit rapport.

Op grond van de eerste aanlevering zijn referentiedatasets gecreëerd met 282 zorgaanbieders voor de CQI VV, 225 zorgaanbieders voor de CQI PG en 135 zorgaanbieders voor de CQI ZT. De kenmerken van de referentiedatasets worden omwille van beknoptheid gepresenteerd in de resultatensectie, tezamen met een vergelijking van de referentiedataset met de zorgaanbieders die niet in de referentiedataset zitten.

### 2.2.2 Analyse, casemixcorrectie en sterrenberekening

Figuur 2.1 geeft een schematisch overzicht van de procedure waarmee de casemixcorrectie geautomatiseerd kan plaatsvinden.

Figuur 2.1 Overzicht van de procedure

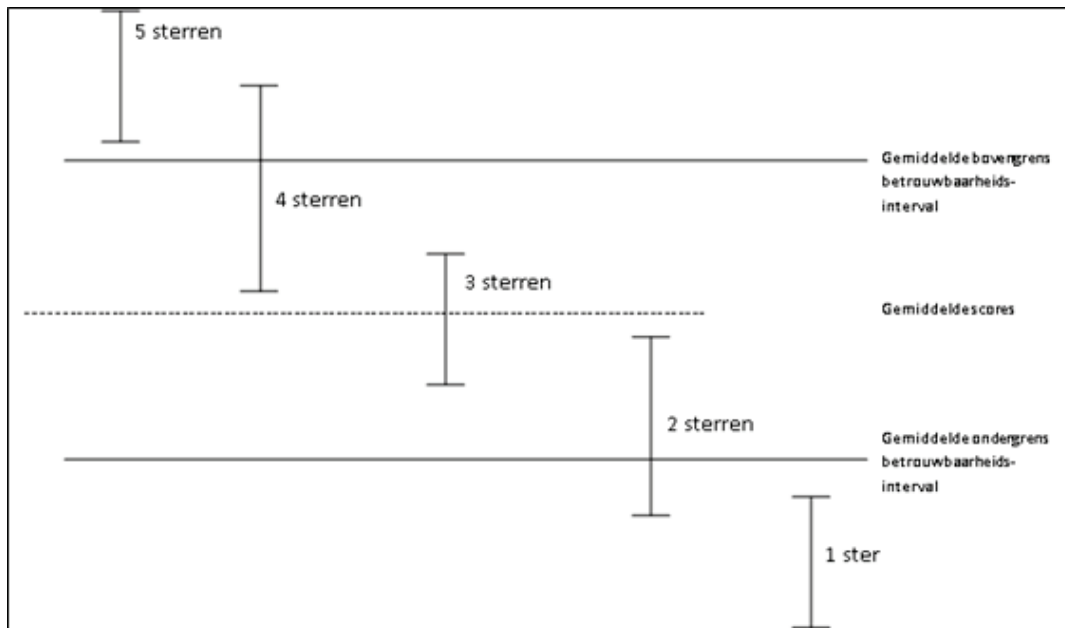


De analyses vinden plaats bij de stap ‘Multilevel analyse en casemix correctie’. Bij deze stap worden multilevel regressie analyses uitgevoerd met twee niveaus: zorgaanbieder en cliënt. De indicatoren vormden de afhankelijke variabelen in de analyses en casemix correctie. Bij de CQI VV werd gecorrigeerd voor leeftijd, gezondheid, opleiding en zorgduur, bij de CQI PG voor leeftijd vertegenwoordiger, opleiding vertegenwoordiger, gezondheid vertegenwoordiger, zorgduur en soort vertegenwoordiger en bij de CQI ZT voor leeftijd, opleiding, gezondheid, zorgduur en hulp van eigen zorgverlener bij het invullen van de vragenlijst. Deze variabelen zijn gerelateerd aan de resultaten op de CQ-index en vanuit de gedachte dat het hier gaat om verschillen in antwoordtendentie tussen demografische groepen waar de zorgaanbieder geen invloed op heeft. Wanneer bijvoorbeeld ouderen minder kritisch zijn, wordt hiervoor gecorrigeerd (Koopman et al., 2011; Wiegiers et al., 2007). Leeftijden werden gecentreerd rond het gemiddelde van de referentiedataset (voor een toelichting op centreren in multilevel modellen, zie Enders en

Tofghi, 2007). Categorische variabelen zijn opgenomen als dummy-variabelen waarbij de grootste groep steeds als referentiegroep gold.

De analyse en casemix correctie resulteerden in empirical Bayes schattingen per instelling (De Boer et al., 2010), voorzien van een vergelijkingsinterval dat is berekend als de schatting  $\pm 1,39$  SE (Goldstein en Healy, 1995). Aan deze schattingen met vergelijkingsinterval werden voor de controle op de resultaten zogeheten sterren toegekend. In figuur 2.2 is visueel weergegeven op welke wijze sterren werden toegekend. Als het interval geheel boven het gemiddelde van de bovengrenzen van alle vergelijkings-intervallen viel, kreeg de zorgaanbieder voor die score vijf sterren. Viel het interval volledig onder het gemiddelde van alle ondergrenzen, dan kreeg de zorgaanbieder één ster. Zorgaanbieders die niet in aanmerking kwamen voor één of vijf sterren kregen vier sterren als hun ondergrens boven het gemiddelde van alle zorgaanbieders lag, of twee sterren als de bovengrens daaronder lag. Overlaptte het interval met het gemiddelde over alle zorgaanbieders, dan kreeg de aanbieder drie sterren (zie figuur 2.2).

Figuur 2.2 Grafische weergave van de wijze waarop sterren worden toegekend. Overgenomen uit Handle with Care (ActiZ, BTN, LOC & ZN, 2013)



Voor interne verbetering worden deze sterren minder bruikbaar geacht en in de spiegelrapportage van ActiZ wordt dan ook geen gebruik gemaakt van sterren. Echter, de consequenties van verschillende keuzes bij dataverwerking en analyse worden met name kritisch bekeken als de resultaten worden gebruikt voor externe verantwoording. In dit licht besteedt dit rapport dan ook de nodige aandacht aan sterren als maat om de consequenties van verschillende keuzes bij dataverwerking en analyses mee te illustreren.

Een belangrijke vraag is welke parameters in de analyse en casemixcorrectie nu één keer worden vastgesteld op grond van de referentiedataset én welke parameters iedere keer opnieuw worden geschat wanneer data van een nieuwe zorgaanbieder wordt toegevoegd. Idealiter wordt alles zoveel mogelijk constant gehouden. Dit is betrekkelijk eenvoudig voor de centreerwaarden en voor de criteria voor het toekennen van sterren (het gemiddelde over alle zorgaanbieder en de gemiddelde onder- en bovengrens). Voor de parameters die het statistiek pakket zelf vaststelt in de multilevel analyse en gebruikt in het schattingsproces is dit aanmerkelijk complexer (regressiecoëfficiënten, varianties en andere parameters die worden gebruikt bij berekening van empirical Bayes schattingen). Deze parameters zijn binnen het statistiek pakket nauwelijks te beïnvloeden en het nabouwen en aanpassen van de algoritmes uit het statistiek pakket is onbegonnen werk met een hoog risico op fouten. Dergelijke parameters zijn dan ook steeds opnieuw vastgesteld waarbij overigens wel is overwogen dat de invloed van één nieuwe zorgaanbieder op deze parameters ook verwaarloosbaar moet zijn gegeven de omvang van de referentiedatasets.

De meeste statistiek pakketten beschikken over de mogelijkheden om het hier beschreven proces volledig in een syntax vast te leggen, zodanig dat het draaien van de syntax erin voorziet dat nieuwe data van een aanbieder tijdelijk worden samengevoegd met de referentiedataset, alle analyses worden uitgevoerd, sterren worden toegekend en de resultaten worden weggeschreven in een resultaatbestand waar alleen de resultaten van die ene zorgaanbieder in behouden blijven. Door deze syntax te koppelen aan een portal waar aanbieders nieuwe data kunnen achterlaten en resultaten kunnen ophalen kan dit proces dagelijks draaien zonder tussenkomst van statistici of onderzoekers. Aangezien de auteurs vooralsnog prijs stelden op een bescheiden visuele controle van resultaten zijn deze controles tot dusver tweewekelijks toegepast waarna de resultaten zijn vrijgegeven.

### **2.2.3 *Representativiteit van referentiedatasets***

Zoals gezegd is het vanuit methodisch oogpunt de koninklijke weg om multilevel vergelijkende analyses in één keer te doen voor alle zorgaanbieders in een sector. Dit garandeert namelijk dat alle parameters in de analyses die van invloed zijn op de resultaten (landelijke gemiddelden, regressiecoëfficiënten, varianties et cetera) zo goed mogelijk worden geschat. Bij analyses op een deel van de zorgaanbieders uit een sector is het mogelijk dat dit deel niet representatief is voor alle zorgaanbieders in de sector en dat dit van invloed is op de resultaten. Tot op zekere hoogte is dit niet geheel te vermijden omdat het niet haalbaar of wenselijk is om te wachten met analyseren totdat alle aanbieders hebben gemeten en aangeleverd. Niettemin is inzicht in de representativiteit van de referentiedatasets van belang. Dit is op verschillende manieren bekeken:

- 1 De kenmerken van respondenten zijn vergeleken tussen de referentiedatasets en de resterende data met behulp van t-toetsen en Chi-kwadraat toetsen.
- 2 De gemiddelden van de ruwe indicatorscores van de referentiedataset zijn vergeleken met gemiddelde indicatorscores van de resterende data met behulp van t-toetsen.
- 3 De sterrenclassificatie die op grond van de geautomatiseerde procedure tot stand is gekomen is vergeleken met de sterrenclassificatie op grond van een analyse van de totale dataset. Het percentage overeenkomst wordt gerapporteerd.

Statistische toetsen gebaseerd op vele duizenden waarnemingen voldoen al snel aan het criterium  $p < 0,05$ , terwijl er misschien niet direct sprake is van noemenswaardige effect. Teneinde dit enigszins te ondervangen is het criterium voor de p-waarde verscherpt van 0,05 tot 0,01 voor de chi-kwadraat toetsen en de t-toetsen.

Aangezien de totale datasets voor meer dan de helft uit referentiedata bestaan is het niet direct verwonderlijk wanneer verschillen in sterrenclassificatie tussen de referentiedataset en de totale dataset (verwaarloosbaar) klein zouden zijn. Om deze reden zijn aanvullende analyses uitgevoerd. Deze analyses bestonden eruit dat er 100 keer een steekproef is getrokken van 100 zorgaanbieders, 100 keer een steekproef van 50 zorgaanbieders, 100 keer een steekproef van 35 zorgaanbieders en 100 keer een steekproef van 20 zorgaanbieders. Voor iedere steekproef is bekeken in hoeverre een analyse van die steekproef leidt tot een andere sterrenclassificatie vergeleken met een het aantal sterren dat die zorgaanbieders zouden krijgen op grond van een analyse van de totale dataset. Dit geeft inzicht in de mate waarin de steekproeffouten die kunnen voorkomen in kleinere referentiedatasets van invloed zijn op de sterrenclassificatie van zorgaanbieders.





## 3 Resultaten

### 3.1 Kenmerken referentiedata versus overige data

In tabel 3.1 staan de kenmerken van de referentiedataset (7.487 cliënten verdeeld over 282 zorgaanbieders) en de overige data (4.273 cliënten verdeeld over 161 zorgaanbieders) weergegeven voor de vragenlijst voor cliënten (CQ-index VV). De verschillen tussen cliënten uit de referentiedata en de overige data zijn gering en alleen voor het opleidingsniveau significant te noemen. De cliënten uit de referentiedataset zijn wat minder hoog opgeleid ( $\chi^2 = 78.83$ ,  $df = 2$ ,  $p < 0,001$ ; zie tabel 3.1).

In tabel 3.2 zijn de kenmerken voor de referentiedataset (6.066 cliënten verspreid over 225 zorgaanbieders) en de overige data (3.059 cliënten verspreid over 131 zorgaanbieders) weergegeven voor vertegenwoordigers van cliënten (CQ-index PG). De verschillen tussen referentiedata en overige data zijn wederom klein en niet significant.

Tabel 3.1 Kenmerken van zorgaanbieders voor interviews met cliënten VV

	referentiedataset	overig	p*
<b>Aantal zorgaanbieders</b>	282	161	
<b>N cliënten</b>	7487	4273	
<b>Leeftijd</b>	85,0	84,6	0,043
<b>Opleiding:</b>			<b>&lt;0,001</b>
laag	72,8%	66,2%	
gemiddeld	23,3%	27,0%	
hoog	3,9%	6,8%	
<b>Gezondheid:</b>			0,093
slecht/matig	56,5%	54,6%	
goed	38,7%	40,7%	
zeer goed/uitstekend	4,8%	4,7%	
<b>Zorgduur:</b>			0,018
tot 6 maanden	8,7%	10,5%	
6 tot 12 maanden	14,9%	15,1%	
1 tot 2 jaar	20,6%	20,5%	
2 tot 5 jaar	32,1%	30,2%	
meer dan 5 jaar	23,6%	23,8%	

\*  $p < 0,01$  wordt als significant beschouwd

Tabel 3.2 Kenmerken van zorgaanbieders voor vertegenwoordigers van cliënten PG

	referentiedataset	overig	p*
<b>Aantal zorgaanbieder</b>	225	131	
<b>N cliënten</b>	6066	3059	
<b>Leeftijd vertegenwoordiger</b>	60,2	60,0	0,533
<b>Opleiding vertegenwoordiger:</b>			0,013
laag	19,1%	17,6%	
gemiddeld	51,1%	49,8%	
hoog	29,7%	32,6%	
<b>Gezondheid vertegenwoordiger:</b>			0,039
slecht/matig	13,3%	13,0%	
goed	59,8%	57,6%	
zeer goed/uitstekend	26,9%	29,5%	
<b>Zorgduur:</b>			0,79
tot 6 maanden	10,2%	9,4%	
6 tot 12 maanden	15,7%	15,8%	
1 tot 2 jaar	23,6%	24,4%	
2 tot 5 jaar	34,3%	34,1%	
meer dan 5 jaar	16,2%	16,3%	
<b>Soort vertegenwoordiger:</b>			0,017
partner/echtgeno(o)t(e)	13,0%	12,8%	
zoon/dochter	70,1%	69,1%	
familie anders	13,4%	13,3%	
geen familie	3,5%	4,9%	

\* p<0,01 wordt als significant beschouwd

De kenmerken voor de referentiedata (5.101 cliënten verspreid over 135 zorgaanbieders) en de overige data (2.574 cliënten verspreid over 56 zorgaanbieders) voor de zorg thuis (CQ-index ZT) staan weergegeven in tabel 3.3 Bij de zorg thuis bleken cliënten uit de referentiedataset gemiddeld bijna één jaar ouder dan de cliënten die niet in de referentiedataset zaten ( $p<0,001$ ). Tevens waren cliënten uit de referentiedataset wat minder hoog opgeleid ( $\chi^2 = 9.53$ ,  $df = 2$ ,  $p=0,009$ ).

Kort samengevat zijn enkele kleine maar significante verschillen gevonden in de individuele kenmerken van de referentiedataset en de overige data. Dit suggereert dat de referentiedatasets wellicht niet geheel representatief zijn voor deze kenmerken. Voor de kenmerken uit de tabellen 3.1 t/m 3.3 wordt echter gecorrigeerd in de casemixcorrectie. Het is dus de vraag of de significante verschillen in kenmerken tussen de referentiedataset en de resterende data nu zo relevant zijn voor de scores van aanbieders. Hier komen we in paragraaf 3.3 op terug.

Tabel 3.3 Kenmerken van zorgaanbieders voor cliënten ZT

	referentiedataset	overig	p*
<b>Aantal zorgaanbieders</b>	135	56	
<b>N cliënten</b>	5101	2574	
<b>Leeftijd</b>	81,7	80,8	<0,001
<b>Opleiding:</b>			<b>0,009</b>
laag	66,9%	65,5%	
gemiddeld	27,6%	27,3%	
hoog	5,4%	7,2%	
<b>Gezondheid:</b>			0,338
slecht/matig	70,6%	69,8%	
goed	27,5%	27,8%	
zeer goed/uitstekend	1,9%	2,4%	
<b>Zorgduur:</b>			0,296
6 tot 12 maanden	12,7%	12,1%	
1 tot 2 jaar	25,6%	24,1%	
2 tot 5 jaar	33,9%	35,8%	
meer dan 5 jaar	27,8%	28,0%	
<b>Hulp van eigen zorgverlener bij invullen:</b>			0,029
geen hulp	93,4%	92,0%	
hulp	6,6%	8,0%	

\* p<0,01 wordt als significant beschouwd

### 3.2 Indicatorscores referentiedata versus overige data

In de vorige sectie keken we naar verschillen in de kenmerken van respondenten tussen de referentiedataset en de overige data. Dit geeft inzicht in de representativiteit van de referentiedataset met betrekking tot deze kenmerken. Minstens zo belangrijk is de representativiteit van de referentiedataset met betrekking tot de ervaren kwaliteit van zorg. Hiertoe zijn de indicatorscores van de referentiedatasets vergeleken met die van de overige data. Tabel 3.4 laat zien dat twee van twaalf indicatoren significant verschilden tussen de referentiedata en de overige data voor de cliënten van verpleeg- en verzorgingshuizen (CQI VV). Voor vertegenwoordigers van cliënten met psychogeriatrische problemen (CQI PG) bleken drie van elf indicatoren significant te verschillen tussen de referentiedata en de overige data. Voor de zorg thuis (CQI ZT) bleken er geen significante verschillen tussen de referentiedataset en de overige data. Er zijn dus enkele significante verschillen tussen de referentiedata en de overige data; in die gevallen waren de scores van de referentiedata (iets) hoger (zie tabel 3.4).

Tabel 3.4 Verschil in indicatorscores tussen referentiedataset en overige data

	interviews	vertegenwoordigers	zorg thuis
# indicatoren	12	11	5
# sign. verschillend (p<0,01)	2	3	0
# referentiedata hogere score	2	3	0

### 3.3 Invloed van de referentiedataset op de sterrenclassificatie

De vergelijking van de resultaten van de geautomatiseerde procedure versus een analyse van de totale dataset is te vinden in tabel 3.5. Te zien is dat de procedure in  $\pm 97\%$  van de gevallen leidt tot hetzelfde aantal sterren als een analyse van de totale dataset. De hierboven beschreven afwijkingen tussen de referentiedataset en de resterende data hebben dus hooguit een beperkte invloed op de sterindeling.

Tabel 3.5 Aantal sterren op grond van analyse van de totale dataset versus de geautomatiseerde procedure voor alle indicatoren

	procedure < totaal	gelijk	procedure > totaal
VV	2,2%	97,0%	0,9%
PG	2,6%	96,8%	0,5%
ZT	0,3%	96,9%	2,8%

### 3.4 Kleinere referentiedatasets

In tabel 3.6 is weergegeven in hoeverre analyses op steekproeven van 20, 35, 50 of 100 zorgaanbieders leiden tot een andere sterrenclassificatie voor die zorgaanbieders vergeleken met de situatie dat die zorgaanbieders worden meegenomen in een analyse van de totale dataset. Het komt geregeld voor dat zorgaanbieders uit een steekproef 1 ster meer of minder krijgen dan wanneer diezelfde zorgaanbieders in een analyse van de totale dataset worden beoordeeld. Bij steekproeven van 20 of 35 zorgaanbieders komen verschillen van 2 sterren nog sporadisch voor, bij steekproeven van 50 zorgaanbieders of meer niet meer. Daarnaast komt het überhaupt niet voor dat het verschil tussen de sterindeling op basis van analyses van een steekproef van zorgaanbieders en een analyse van de totale dataset meer dan twee sterren bedraagt.

Tabel 3.6 Verschil in sterindeling tussen analyses van steekproeven van 20, 35, 50 en 100 zorgaanbieders versus de sterindeling voor die zorgaanbieders op grond van een analyse van de gehele dataset (gemiddelden van 100 steekproeven).

	sterindeling op basis van analyses totale dataset minus sterindeling op basis van een steekproef*							
	#zorgaanbieders steekproef/totaal	- 2	- 1	hetzelfde	+ 1	+ 2	5 <sup>e</sup> percentiel verschil_0	95 <sup>e</sup> percentiel verschil_0
<b>Interviews met cliënten VV (443 zorgaanbieders in totaal):</b>								
20 zorgaanbieders	5%	0,016%	8,120%	85,244%	6,599%	0,020%	80,083%	90,486%
35 zorgaanbieders	8%	0,005%	6,736%	88,672%	4,585%	0,002%	84,891%	92,563%
50 zorgaanbieders	11%	0,000%	4,493%	90,702%	4,805%	0,000%	87,323%	93,915%
100 zorgaanbieders	23%	0,000%	3,850%	93,505%	2,644%	0,000%	91,060%	95,334%
<b>Vertegenwoordigers van cliënten PG (356 zorgaanbieders in totaal):</b>								
20 zorgaanbieders	6%	0,000%	6,517%	87,763%	5,721%	0,000%	80,998%	93,123%
35 zorgaanbieders	10%	0,000%	5,898%	90,964%	3,138%	0,000%	86,777%	94,671%
50 zorgaanbieders	14%	0,000%	4,970%	92,374%	2,657%	0,000%	87,178%	95,332%
100 zorgaanbieders	28%	0,000%	3,226%	95,009%	1,765%	0,000%	92,260%	96,892%
<b>Cliënten ZT (191 zorgaanbieders in totaal):</b>								
20 zorgaanbieders	10%	0,000%	3,336%	89,998%	6,666%	0,000%	81,026%	96,000%
35 zorgaanbieders	18%	0,000%	2,583%	92,240%	5,177%	0,000%	86,931%	96,460%
50 zorgaanbieders	26%	0,000%	1,770%	93,330%	4,900%	0,000%	89,167%	96,909%
100 zorgaanbieders	52%	0,000%	0,911%	95,218%	3,871%	0,000%	92,310%	97,510%

\* verschillen groter dan plus of min twee sterren kwamen niet voor

In tabel 3.6 is eveneens te zien dat naarmate er meer zorgaanbieders in de steekproef zitten, de overeenkomst toeneemt tussen de resultaten van de steekproef en die van een analyse op de totale dataset. Dit is te zien bij interviews met cliënten VV, vertegenwoordigers van cliënten PG én de cliënten ZT. Voor de zorg thuis is de overeenkomst voor een gegeven aantal zorgaanbieders daarnaast iets groter. Dit heeft er vermoedelijk mee te maken dat de totale dataset voor de zorg thuis minder zorgaanbieders bevat, waardoor een steekproef van bijvoorbeeld 20 zorgaanbieders een groter deel van de totale dataset dekt.

De rechterkolommen uit tabel 3.6 geven voor iedere serie van 100 steekproeven steeds het 5<sup>e</sup> en het 95<sup>e</sup> percentiel weer van de overeenkomst tussen resultaten van de steekproef en die van een analyse van de totale dataset. Dit illustreert dat hoewel de sterindeling van een steekproef van 20 zorgaanbieders voor cliënten van verpleeg- en verzorgingshuizen *gemiddeld* voor 85% overeenkomt met die van een analyse van de totale dataset, dit in ongunstige gevallen kan dalen tot minder dan 81% (zie tabel 3.6).

## 4 Discussie

### 4.1 Kernbevindingen

In dit rapport is een strategie gepresenteerd voor het versnellen van vergelijkende analyses en casemixcorrectie met behulp van multilevel analyses door gebruik van een referentiedataset. Deze strategie is inmiddels meer dan een jaar operationeel in de VVT en draagt ertoe bij dat zorgaanbieders die een spiegelrapportage aanvragen bij ActiZ binnen enkele weken beschikken over casemixgecorrigeerde resultaten die zijn berekend volgens de daartoe aangewezen methoden. Hiermee zijn deze resultaten veel sneller en beter inzetbaar voor interne verbetering.

In combinatie met een portal is de procedure volledig te automatiseren, zodanig dat binnen enkele uren na het aanleveren van nieuwe data aan de portal, de resultaten klaar staan in diezelfde portal. In het huidige project is er echter voor gekozen resultaten vrij te geven na een tweewekelijks controle teneinde zicht te houden op eventuele onwaarschijnlijke of verdachte resultaten, die zich overigens niet voor hebben gedaan. Het uitgangspunt voor de procedure was de gedachte dat een zorgaanbieder die – om wat voor reden dan ook – meerdere keren exact dezelfde data aanbiedt ook exact dezelfde resultaten moet terugkrijgen. Hiertoe zijn de parameters die een rol spelen in de analyses waar mogelijk constant gehouden. Voor verschillende parameters was het echter niet goed mogelijk om deze een vaste waarde te geven, omdat die waarden geschat worden binnen het analyseproces en direct worden meegenomen in verschillende algoritmes. Om dit op te lossen is gebruik gemaakt van een vaste referentiedataset waarmee de ‘input’ voor de analyses constant wordt gehouden. Dit zorgt ervoor dat parameters die van invloed zijn op de resultaten en buiten de controle van de onderzoeker of statisticus liggen, toch iedere keer dezelfde waarde krijgen.

Een voor de hand liggende vraag bij het gebruik van een referentiedataset is of deze referentiedataset representatief is. De referentiedatasets uit dit rapport waren veelal niet (geheel) representatief voor alle beschikbare data. Toch heeft dit nauwelijks invloed gehad op het aantal sterren dat werd toegekend: de procedure leverde in  $\pm 97\%$  van de gevallen exact hetzelfde aantal sterren op als een analyse van de totale dataset. Dit komt deels omdat de referentiedatasets groot waren, aangezien aanvullende analyses bevestigden dat de overeenkomst in sterren tussen een analyse van de steekproef en een analyse van de totale dataset steeg naarmate steekproeven groter waren.

## 4.2 Ervaringen met het geautomatiseerde proces: aandachtspunten

Wanneer ondeugdelijke data worden aangeboden bij een geautomatiseerde procedure leidt dit uiteraard tot ondeugdelijke resultaten. Binnen het huidige traject is het geautomatiseerde proces dan ook steeds vergezeld van enkele checks alvorens de resultaten vrij te geven. Hiermee komen zeer onwaarschijnlijke resultaten aan het licht alvorens zij worden vrijgegeven. Toch blijft het lastig te beoordelen of onwaarschijnlijke resultaten nu een gevolg zijn van ondeugdelijke data of dat het gaat om een opmerkelijke zorgaanbieder. Daarnaast is het ook denkbaar dat ondeugdelijke gegevens leiden tot resultaten die niet direct opmerkelijk zijn, waardoor ook niet direct aan het licht komt dat er iets niet klopt. Dit kan bijvoorbeeld gebeuren als bij een indicator die uit meer vragen bestaat een coderingsfout wordt gemaakt bij één van de onderliggende vragen.

De beste manier om zicht te houden op de kwaliteit van de data is om – naast de standaard checks en validaties op bestandslay-out, onmogelijk waarden et cetera – met enige regelmaat aanvullende checks te doen op bijvoorbeeld systematische verschillen tussen meetbureaus, opmerkelijke en onverwachte trends in scores et cetera. Een belangrijk aandachtspunt blijft niettemin dat de partij die data aanbiedt primair verantwoordelijk is voor de kwaliteit van die data. Aangezien volledig waterdichte controleprocedures niet of nauwelijks te realiseren zijn, is het raadzaam de primaire verantwoordelijkheid van de aanbieder partij duidelijk te blijven benadrukken in alle communicatie.

## 4.3 Referentiedatasets: eisen en onderhoud

Uit het huidige rapport is gebleken dat referentiedatasets leiden tot resultaten die (enigszins) afwijken van de resultaten van één analyse van de totale dataset. Dit is *onwenselijk* vanuit de gedachte dat een analyse van de totale dataset de werkelijkheid beter weerspiegelt. Het is echter *niet oneerlijk* omdat alle zorgaanbieders in exact dezelfde procedure zijn geanalyseerd met exact dezelfde referentiedataset. De spelregels zijn voor alle zorgaanbieders dus gelijk. Daarnaast is een analyse achteraf op basis van een totale dataset geregeld niet haalbaar wanneer dit leidt tot een onverantwoord lange wachttijd voor zorgaanbieders en cliënten die als eerste hun metingen voltooien.

Omdat afwijkingen tussen analyses op basis van een referentiedataset en analyses op basis van de totale dataset onwenselijk zijn, is het van belang om enige eisen te stellen aan de referentiedataset. Wat betreft de omvang van de referentiedataset gaat het hierbij niet alleen om het aantal zorgaanbieders, maar ook om het aantal zorgaanbieders in de totale populatie; een referentiedataset van 50 zorgaanbieders zal immers representatiever zijn voor een populatie van 60 zorgaanbieders dan voor een populatie van 200 zorgaanbieders.

Uit dit rapport is gebleken dat bij steekproeven van 50 of meer zorgaanbieders (17 – 37 % van de totale dataset), de sterindeling gemiddeld voor meer dan 90% overeenkomt met de indeling die deze zorgaanbieders zouden krijgen op basis van een analyse van de totale



dataset. Wanneer de zorgaanbieders uit de steekproef minder dan 10% van de zorgaanbieders uit de totale dataset beslaan daalt dit tot onder de 90%. De vraag blijft niettemin in welke mate resultaten van een referentiedataset overeen moeten stemmen met die van een analyse op een totale dataset achteraf. Aangezien hier verschillend over gedacht kan worden is het van belang om bij het werken met een referentiedataset in overleg met belanghebbenden en experts tot consensus te komen over welke streefwaarden voor overeenstemming gewenst en haalbaar zijn. Uiteraard dient te worden gemonitord in hoeverre die streefwaarden in de praktijk ook worden gehaald. In afwachting van meer evidentie en/of consensus hierover lijkt het redelijk om voor een referentiedataset minimaal 30 zorgaanbieders te nemen (conform de aanbeveling uit het CQI Handboek voor multilevel analyses [Koopman et al., 2013]) en zo nodig aan te vullen tot 15–20% van de markt.

Tot slot is het de vraag met welke frequentie de referentiedataset wordt geactualiseerd. Het is immers zeer goed denkbaar dat een gegeven referentiedataset op zeker moment steeds minder representatief wordt omdat de kwaliteit van zorg binnen een sector verandert of omdat de samenstelling van cliënt-/patiëntenpopulaties verandert. Omdat dit soort veranderingen zich vermoedelijk over enkele jaren uitstrekken lijkt een actualisatie van de referentiedataset iedere één a twee jaar toereikend. Echter, binnen die termijn worden één of meer indicatoren vaak ook aangescherpt, wat op zichzelf ook weer een nieuwe referentiedataset vereist. Het actualiseren of vernieuwen van referentiedatasets zou in de praktijk dus vaak samen kunnen vallen met aanpassingen aan de indicatoren.

#### **4.4 Conclusie**

Dit rapport heeft laten zien dat het automatiseren van multilevel vergelijkende analyses met behulp van een referentiedataset mogelijk is. De resultaten van de versnelde procedure komen nauw overeen met die van een analyse van de totale dataset achteraf als de referentiedataset groot is. Naarmate de referentiedataset kleiner is zal dit minder het geval zijn. Partijen die op vergelijkbare wijze ook een versnelde en geautomatiseerde procedure willen inrichten doen er verstandig aan de algemene aanbevelingen voor het aantal zorgaanbieders in een multilevel analyse te volgen én te streven naar een referentiedataset met een substantieel deel van de zorgaanbieders in een sector. Vanuit pragmatische overwegingen ligt het daarbij voor de hand zo'n procedure in te richten als de indicatoren al enigszins zijn uitgekristalliseerd, om te voorkomen dat er steeds weer nieuwe referentiedatasets nodig zijn als gevolg van wijzigingen in de indicatoren.

Het versnellen en automatiseren van multilevel analyses draagt bij aan een snellere terugkoppeling van resultaten van metingen. Dit draagt bij aan de bruikbaarheid van resultaten omdat deze actueler zijn als ze beschikbaar komen. Teneinde kennis en consensus op te bouwen over (vereiste) overeenstemming tussen continue geautomatiseerde procedures en een eenmalige analyse achteraf is het van belang om deze overeenstemming te blijven monitoren voor verschillende sectoren en typen indicatoren.



# Literatuur

- ActiZ, BTN, LOC & ZN. 'Handle with Care'. Een handreiking voor het gebruik van de resultaten van het kwaliteitsdocument VVT 2012. Cliëntgebonden indicatoren. Versie 2.0. Utrecht: ActiZ, 2013
- 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(3Pt1): 1177-99
- Boer D de, Hoek L van der, Delnoij D, Groenewegen P. *Kleine zorgaanbieders in multilevel vergelijkende analyses: de CQI verpleging, verzorging en thuiszorg*. Utrecht: NIVEL, 2010
- <http://www.centrumklantervaringzorg.nl/cqi-richtlijnen/handboek-eisen-en-werkwijzen-cqi-metingen.html>. *Handboek Eisen en Werkwijzen CQI-metingen 2011*. 7/26/2013
- [http://www.centrumklantervaringzorg.nl/vragenlijsten/verpleging-verzorging-en-thuiszorg/verpleging-verzorging-en-thuiszorg.html](http://www.centrumklantervaringzorg.nl/vragenlijsten/verpleging-verzorging-en-thuiszorg/verpleging-verzorging-en-thuiszorg/verpleging-verzorging-en-thuiszorg.html). *Vragenlijsten en werkinstructies 2012*
- Damman OC, Stubbe JH, Hendriks M, Arah OA, Spreuwenberg P, et al. Using multilevel modeling to assess case-mix adjusters in consumer experience surveys in health care. *Med Care*, 2009; 47(4):496-503
- Enders CK, Tofighi D. Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychol Methods*, 2007;12(2):121-38
- Goldstein H, Healy MJR. The graphical presentation of a collection of means. *J Royal Stat Soc Seris A*, 1995;158(1):175-7
- Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance. *J Royal Stat Soc Seris A*, 1996;159(3):385-443
- Greenland S. Principles of multilevel modelling. *Int J Epidemiol*, 2000;29(1):158-67
- Hopman P, Boer D de, Rademakers J. *Kennisvraag: wat heeft vijf jaar CQ-index opgeleverd?* Utrecht: NIVEL, 2011
- Inspectie voor de Gezondheidszorg, LOC Zeggenschap in zorg, Zorgverzekeraars Nederland *Kwaliteitsdocument VV&T*. Utrecht: IGZ, 2013
- Koolman X, Visser J, Appelman M. Indicatorstandaard 1.5. *Programma Zichtbare Zorg*, 2012

- Koopman L, Sixma H, Boer D de, Delnoij D. *Handboek CQI Ontwikkeling. Richtlijnen en voorschriften voor de ontwikkeling van een CQI meetinstrument*. Utrecht: Centrum Klantervaring Zorg, 2011
- Koopman L, Sixma H, Boer D de, Delnoij D. *Handboek eisen en werkwijzen CQI metingen*. Utrecht: Centrum Klantervaring Zorg, 2013
- Ministerie van Volksgezondheid, Welzijn en Sport. *Van systemen naar mensen*. Beleidsbrief. Den Haag: Ministerie van VWS, 2013
- Sixma H, Hendriks M, Boer D de, Delnoij D. *Handboek CQI Metingen: richtlijnen en voorschriften voor metingen met een CQI meetinstrument 2008*. Utrecht: NIVEL/CKZ, 2008
- Snijders TAB, Bosker RJ. *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. Londen: Sage Publishers, 1999
- 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