Methodologische reeks

Methode imputatie LBZ-gegevens

Marjolein Peters
Corine Penning
Henrico Witvliet
Ton de Waal
Agnes de Bruin

Oktober 2018

Inhoudsopgave

1. Inleiding

De statistieken van het CBS over ziekenhuisopnamen naar diagnose zijn gebaseerd op de ziekenhuiszorgregistraties van DHD (Utrecht). Vanaf verslagjaar 2013 is de Landelijke Medische Registratie (LMR) van DHD overgegaan in de Landelijke Basisregistratie Ziekenhuiszorg (LBZ), met een nieuw datamodel. Voor het CBS was dit aanleiding om de processen en statistische producten gebaseerd op de LMR te herontwerpen, met de LBZ-data als uitgangspunt.

Een belangrijke verandering in de LBZ is dat ziekenhuizen die niet of niet volledig LBZ-gegevens geregistreerd hebben, nu geacht worden van de missende opnamen wel LBZ-microrecords aan te leveren met een aantal basale persoons- en opnamegegevens die direct uit de ziekenhuisadministratie komen. Voorheen waren van deze opnamen alleen enkele randtotalen per ziekenhuis bekend, op basis waarvan voor de missende opnamen records werden 'gegenereerd' door de registratiehouder. Deze bijschatting werd gedaan door het aanmaken van duplicaatrecords uit de wel geregistreerde opnamen met dezelfde basiskenmerken (specialisme, soort opname en woongemeente) als de missende opnamen. Maar vanaf 2013 zijn van de voorheen missende opnamen dus werkelijke (maar incomplete) microrecords beschikbaar in de LBZ, die in principe ook koppelbaar zijn met de Basisregistratie Personen (BRP).

Bij deze incomplete records mist er onder andere informatie over diagnosen. Voor de ziekenhuiszorgstatistieken van het CBS is het daarom belangrijk om de missende variabelen in deze records aan te vullen. Het CBS heeft hiervoor een imputatieprocedure ontwikkeld. Met deze imputatie wordt de missende informatie bij de incomplete records opgevuld; dat wil zeggen, we schatten de waardes voor de niet geregistreerde variabelen en vullen het bestand aan met deze geschatte waardes.

De imputatieprocedure zoekt naar een complete opname die zoveel mogelijk overeenkomt met de incomplete opname en vervolgens worden de ontbrekende gegevens van de incomplete opname gevuld met die van de gematchte complete opname. Hierbij wordt iedere opname als een afzonderlijk event beschouwd. Er wordt dus niet gekeken of er meerdere opnamen waren bij dezelfde persoon. Een persoon die meerdere keren in het ziekenhuis is opgenomen krijgt dus niet noodzakelijkerwijs dezelfde diagnose toegewezen bij deze opnamen. Hier is voor gekozen omdat gebleken is dat relatief weinig patiënten meerdere keren per jaar met dezelfde diagnose in het ziekenhuis worden opgenomen. Ook wordt niet specifiek gezocht naar opnamen binnen hetzelfde ziekenhuis; er wordt alleen rekening gehouden met het soort ziekenhuis. Dit is gedaan omdat de groepen waarbinnen de matchende opnamen gezocht worden voldoende groot moeten zijn. Als er variabelen ontbreken bij een specifieke groep (b.v. een bepaald specialisme of zorgtype) in een bepaald ziekenhuis, dan is dit vaak voor alle opnamen in die groep het geval en zijn er dus binnen hetzelfde ziekenhuis onvoldoende matchende opnamen te vinden. Met de imputatieprocedure willen we dus zo goed mogelijk matchende complete opnamen vinden voor de incomplete opnamen, met als doel het maken van populatieschattingen (met name naar diagnose).

De imputatieprocedure is ontwikkeld in programmeertaal R (versie 3.3.2) *[1]* en bij de ontwikkeling is gebruik gemaakt van LBZ-gegevens van 2013. Deze dataset bevat 3,9 miljoen ziekenhuisopnamen: dit zijn ongeveer 3 miljoen complete opnamen en zo'n 960.000 incomplete opnamen (waarvan 280.000 klinische opnamen en 680.000 dagopnamen). De kwaliteit van de

imputatieprocedure is getest met de complete LBZ-opnamen van 2015; dit betrof 3,1 miljoen ziekenhuisopnamen.

In deze notitie beschrijven we de ontwikkelde imputatieprocedure (paragraaf 2). De kwaliteit van de imputatieprocedure wordt beschreven in paragraaf 3; bij de kwaliteitsschatting is alleen gekeken naar een selectie van complete records die we hebben geïmputeerd. In werkelijkheid is de invloed van de imputatie op de uitkomsten natuurlijk een stuk minder groot, omdat dan ook de uitkomsten van alle complete records worden meegenomen. In paragraaf 4 besluit deze notitie met enkele conclusies over de imputatie.

2. Imputatiemethode

2.1 Imputatiemethoden in het algemeen

Voor het imputeren van ontbrekende data bestaan diverse methoden. Een deel van deze methoden is gebaseerd op expliciete modellen voor de complete maar deels onbekende data. Deze methoden schatten dan allereerst de modelparameters. De geschatte modellen worden vervolgens gebruikt om imputatiewaardes voor de ontbrekende data te bepalen. Een theoretisch nadeel van deze methoden is dat ze foutieve resultaten kunnen geven als het model incorrect is. Een belangrijk praktisch nadeel is dat ze vaak behoorlijk wat rekentijd vergen, zeker voor een groot databestand als de LBZ. Vanwege dit laatste nadeel zijn imputatiemethoden gebaseerd op expliciete modellen niet geschikt voor de LBZ-data. We zullen verder dan ook niet meer ingaan op dergelijke methoden.
Andere methoden maken geen gebruik van expliciete modellen, maar gaan rechtstreeks uit van de waargenomen data. Dit zijn meestal zogeheten hot-deck methoden. Bij hot-deck methoden worden ontbrekende waardes aangevuld met waargenomen data uit hetzelfde databestand. Er bestaan verschillende soorten hot-deck methoden. De meest gebruikte hot-deck methoden zijn random hot-deck (in groepen) en nearest-neighbour imputatie. We zullen beide methoden kort bespreken.

2.2 Hot-deck imputatie

2.2.1 Random hot deck (in groepen)

Bij random hot-deck (in groepen) maakt men op basis van achtergrondkenmerken, zoals leeftijd en geslacht, groepen van soortgelijke opnamen. Om ontbrekende waardes voor een opname (de ontvanger) te imputeren wordt dan random een andere opname (de donor) uit dezelfde groep getrokken waarvoor deze waardes wel zijn waargenomen. De waargenomen waardes van de donor worden dan gebruikt om alle ontbrekende waardes van de ontvanger te imputeren. Als de opnamen met ontbrekende waardes op de volledig waargenomen opnamen lijken, dan kan random hot-deck (in groepen) worden gebruikt om totalen voor de variabelen met ontbrekende waardes te schatten. Bovendien blijven relaties tussen de variabelen met ontbrekende waardes en de achtergrondkenmerken die voor het maken van de groepen zijn gebruikt over het algemeen goed behouden. Relaties tussen de variabelen met ontbrekende waardes en achtergrondkenmerken die niet voor het maken van de groepen zijn gebruikt kunnen echter wel verstoord raken.

Een praktisch nadeel van random hot-deck (in groepen) is dat groepen te weinig donorrecords kunnen bevatten als men gebruik wil maken van veel achtergrondkenmerken. In dat geval moeten deze groepen worden samengevoegd met andere groepen om betrouwbare schattingen te krijgen. Dit nadeel kan de toepassing van random hot-deck (in groepen) in de praktijk lastig maken. Bij ieder groepje dat te weinig donorrecords bevat moet een afweging worden gemaakt met welke andere groep dat groepje moet worden samengevat.

2.2.2 Nearest-neighbour imputatie

Bij nearest-neighbour imputatie worden de achtergrondkenmerken gebruikt om een afstand tussen records te berekenen. Voor ieder incompleet record wordt dan de meest dichtstbijzijnde donor gezocht, dat wil zeggen een compleet record dat qua achtergrondkenmerken het beste overeenkomt met de ontvanger. De ontbrekende waardes in de ontvanger worden dan opgevuld met de corresponderende waardes uit de donor. Op deze basisversie bestaan enkele varianten. In een eerste stap kan je, bijvoorbeeld, de *k* meest dichtstbijzijnde donoren selecteren, en vervolgens hier random één uitkiezen om de ontbrekende waardes in de ontvanger daadwerkelijk te imputeren.
Een voordeel van nearest-neighbour imputatie in vergelijking met random hot-deck imputatie (in groepen) is dat op eenvoudige wijze een groot aantal achtergrondkenmerken kan worden meegenomen. Hierdoor blijven meer relaties tussen de variabelen met ontbrekende data en de achtergrondkenmerken behouden. Bovendien treedt bij nearest-neighbour imputatie nooit het probleem op dat groepjes te klein worden, simpelweg omdat er geen groepjes worden gemaakt.
Nadeel van de nearest-neighbour imputatie is dat voor alle combinaties van complete en incomplete records een afstand berekend moet worden. Dit gaat om zeer veel combinaties (in 2013: 3 miljoen complete records x 900.000 incomplete records = 2700 miljard combinaties) en kost daarmee heel veel rekenkracht.

Voor de LBZ hebben we gekozen voor een combinatie van random hot deck en nearest-neighbour imputatie: we gebruiken de afstandsfunctie van de nearest-neighbour imputatie om de best passende donor te kiezen, maar we doen dit in vooraf bepaalde startgroepen (zoals bij random hot deck imputatie). Deze keuze is gemaakt omdat we willen dat bepaalde belangrijke kenmerken altijd overeenkomen, en om de rekentijd van de nearest-neighbour imputatie in te perken. Daarom zijn startgroepen bepaald op basis van die drie achtergrondkenmerken die voor iedere opname in de LBZ bekend zijn en die van groot belang zijn voor de te imputeren diagnosen, namelijk: zorgtype (dagopname, klinische opname of observatie), specialisme (46 categorieën) en geslacht (man/vrouw). In totaal zijn er dus 3 x 46 x 2 = 276 startgroepen. Binnen iedere startgroep (d.w.z. elke combinatie van specialisme, zorgtype en geslacht) passen we de nearest-neighbour imputatie toe, waarbij we dus alleen de records uit die groep gebruiken als ontvangers (incomplete opnamen) én donoren (complete opnamen).

Aangezien het om slechts drie achtergrondkenmerken gaat bij het bepalen van de startgroepen is het praktische probleem van te kleine groepen (zoals beschreven bij random hot deck) met deze gecombineerde methode veel minder groot; in 2013 zijn er slechts 12 startgroepen waarin geen donoren beschikbaar zijn, bijvoorbeeld de klinische opnamen bij het specialisme “Inwendige geneeskunde – subspecialisme Immunologie”; zowel bij mannen als vrouwen waren hier geen donoren. Voor deze situaties is gezocht naar sterk gelijkende groepen waar wel voldoende donoren beschikbaar zijn. In dit voorbeeld is gekozen voor klinische opnamen met het hoofdspecialisme “Inwendige geneeskunde” (incl. alle subspecialismen).

Het in praktijk vaak lastigste aspect van nearest-neightbour imputatie is de keuze van de afstandsfunctie en de inregeling hiervan: hoe zwaar telt het ene achtergrondkenmerk mee in vergelijking met een ander achtergrondkenmerk. Voor de LBZ zijn de mogelijkheden echter zeer beperkt. Dat komt doordat LBZ-data uit een mix van numerieke en categorische variabelen bestaat. Slechts weinig afstandsfuncties kunnen met zo'n mix van numerieke en categorische variabelen omgaan. De zogeheten Gower-afstand is hiervan de bekendste. Deze Gower-afstand is gebruikt voor de LBZ imputatie en wordt in de volgende paragraaf (2.3) besproken.

Binnen iedere startgroep wordt de Gower-afstand gebruikt om bij elke ontvanger de dichtstbijzijnde donor te vinden. De ontbrekende waardes van de ontvanger worden met de waardes van die donor geïmputeerd. De variabelen "specialisme", "zorgtype" en "geslacht" komen altijd overeen bij de ontvanger en de donor (vanwege de groepenindeling vooraf), tenzij er dus geen donoren beschikbaar zijn binnen de groep; dan is het specialisme anders.

2.3 Gower-afstand

De Gower-afstand meet in hoeverre twee records verschillen. De records mogen hierbij uit een mix van dichotome, nominale, ordinale en numerieke variabelen bestaan. Alle variabelen tellen in de Gower-afstand in principe even zwaar mee. De Gower-afstand tussen twee records is een som van de afstanden tussen de afzonderlijke variabelen in deze records. Een Gower-afstand van 0 betekent een perfecte match en een grote Gower-afstand (maximaal 1) duidt op grote verschillen tussen de ontvanger en donor.

De Gower-afstand wordt in formulevorm gegeven door

$$D_G\left(\boldsymbol{x}^{(o)}, \boldsymbol{x}^{(d)}\right) = \frac{1}{p}\sum_{i=1}^{p} D\left(x_i^{(o)}, x_i^{(d)}\right) \quad (1)$$

Hierbij is $\boldsymbol{x}^{(o)}$ een ontvanger-record, $\boldsymbol{x}^{(d)}$ een donor-record, p het aantal variabelen, $x_i^{(o)}$ de waarde van variabele i $(i = 1, \dots, p)$ in het ontvanger-record, $x_i^{(d)}$ de waarde van variabele i in het donor-record, $D\left(x_i^{(o)}, x_i^{(d)}\right)$ de afstand tussen de waardes van variabele i in het ontvanger-record en het donor-record, en $D_G\left(\boldsymbol{x}^{(o)}, \boldsymbol{x}^{(d)}\right)$ de totale Gower-afstand tussen het ontvanger-record en het donor-record.
Voor alle variabelen i $(i = 1, \dots, p)$ ligt de afstand tussen de waardes van variabele i in het ontvanger-record en het donor-record, dat wil zeggen $D\left(x_i^{(o)}, x_i^{(d)}\right)$, tussen 0 en 1. Variabelen die ontbreken in een ontvanger worden niet meegenomen bij de berekening van de Gower-afstand van die ontvanger. Voor die variabelen is $D\left(x_i^{(o)}, x_i^{(d)}\right)$ dus gelijk aan nul.
De Gower-afstand voor records waarbij alle variabelen aanwezig zijn is dus eigenlijk niet vergelijkbaar met de Gower-afstand voor records waarbij bijvoorbeeld twee variabelen missen. Met andere woorden: een "overall" waarde van de Gower-afstand is dus eigenlijk niet te bepalen voor een imputatieronde, als er ontvangers zijn waarbij een deel van de variabelen ontbreekt.

Als de i-de variabele een dichotome of nominale variabele is, dan is

$$D\left(x_i^{(o)}, x_i^{(d)}\right) = 0 \text{ als } x_i^{(o)} = x_i^{(d)} \quad (2)$$

of

$$D\left(x_i^{(o)}, x_i^{(d)}\right) = 1 \text{ als } x_i^{(o)} \neq x_i^{(d)} \quad (3)$$

Als de i-de variabele een numerieke variabele is, dan is

$$D\left(x_i^{(o)}, x_i^{(d)}\right) = \frac{|x_i^{(o)} - x_i^{(d)}|}{R_i} \quad (4)$$

waarbij R_i de range van variabele i is, dat wil zeggen het verschil van de maximale en de minimale waarde van variabele i. Als de range R_i van een variabele i, bijvoorbeeld, 100 is, dan levert een verschil van 1 tussen $x_i^{(o)}$ en $x_i^{(d)}$ dus een bijdrage van 1/100 aan de Gower-afstand.

In principe kan je de Gower-afstand (formule 1) uitbreiden door verschillende gewichten aan de variabelen toe te kennen. Hierdoor telt de ene variabele zwaarder mee dan een andere. Voor meer informatie over de Gower-afstand, zie bijgevoegde referenties *[2-4]*.

Het aantal R-packages waarin de Gower-afstand beschikbaar is, is helaas zeer beperkt. In de wel bekende R-packages waarin de Gower-afstand beschikbaar is bestaan geen eenvoudige mogelijkheden om de achtergrondkenmerken verschillende gewichten in de afstandsfunctie te geven. De enige manier om een achtergrondkenmerk een hoger gewicht te geven is door dit kenmerk een aantal keer te kopiëren en aldus meerdere keren in de afstandsfunctie op te nemen. Deze aanpak hebben we echter niet gevolgd, zodat iedere variabele even zwaar meetelt. In onze implementatie hebben we gebruikgemaakt van *gower.dist* uit het R-package S*tatMatch [5]* met aanpassingen door het CBS om de rekentijd te verkorten. Bovendien hebben we de bijdragen van de numerieke variabelen iets aangepast door de range R_i af te kappen, bijvoorbeeld bij leeftijd: in plaats van ieder jaar de maximale leeftijd te gebruiken hebben we de range afgekapt op 95 jaar. Zo zijn ieder jaar de gewichten hetzelfde en hebben de uitbijters (klein aantal zeer oude mensen) geen invloed op de gewichten per jaar.

2.4 Implementatie van de Gower-afstand in de LBZ imputatie

De Gower-afstand in de LBZ imputatie wordt berekend met de volgende variabelen:

1. Soort ziekenhuis (academisch ziekenhuis, topklinisch algemeen ziekenhuis, of overig algemeen ziekenhuis)
2. Leeftijd in jaren (gemaximeerd op 95 jaar)
3. Migratieachtergrond uit BRP (autochtoon, Marokko, Turkije, Suriname, voormalige Nederlandse Antillen en Aruba, overige niet-westerse landen, overige westerse landen, of migratieachtergrond onbekend)
4. Overlijden in ziekenhuis (ja/nee)
5. Opnameduur in dagen (gemaximeerd op 365 dagen)
6. Urgentie van de opname (acuut/niet acuut)

De variabelen soort ziekenhuis (1), BRP migratieachtergrond (3), overlijden in ziekenhuis (4) en urgentie van de opname (6) zijn nominale variabelen. De variabelen leeftijd (2) en opnameduur (5) zijn numerieke variabelen. Leeftijd en opnameduur zijn gemaximeerd (leeftijd op 95 jaar; opnameduur op 365 dagen) om de maximale afstanden (en daarmee de gewichten) constant te houden bij toepassing van de imputatie over verschillende jaren.

De twee in de LBZ aanwezige categorale ziekenhuizen (een kankerkliniek en een oogziekenhuis) zijn in principe ingedeeld in de categorie ‘topklinisch algemeen ziekenhuis’. Alleen voor de ontvangers met een opname in de kankerkliniek vindt een aparte imputatie plaats (zie paragraaf 2.5).

Voor ieder van de zes variabelen wordt een afstand berekend tussen de ontvanger en alle mogelijke donoren: deze afstanden variëren tussen 0 en 1. Ter illustratie een voorbeeld wat betreft de nominale variabele overlijden in ziekenhuis: iemand kan tijdens een opname wel of niet overlijden. Als zowel donor als ontvanger zijn overleden, dan is de afstand voor deze variabele gelijk aan 0. Als zowel de donor als de ontvanger niet zijn overleden, dan is de afstand ook gelijk aan 0. Als overlijden in het ziekenhuis niet overeen komt tussen donor en ontvanger, dan is de afstand gelijk aan 1. Een tweede voorbeeld is de berekende afstand bij de numerieke variabele leeftijd: als donor en ontvanger 21 jaar in leeftijd verschillen, dan is de afstand voor deze variabele gelijk aan: $\frac{21}{95} = 0{,}22105.$ Als donor en ontvanger slechts 3 jaar in leeftijd verschillen, dan is deze afstand dus gelijk aan $\frac{3}{95} = 0{,}03158.$

De totale Gower-afstand voor een ontvanger-donor combinatie in de LBZ wordt bepaald door de som van de afzonderlijke afstanden voor de zes variabelen te delen door zes (de gemiddelde afstand). Vervolgens wordt de donor gekozen met de kleinste totale afstand tot de ontvanger. Als er meerdere donoren zijn met de kleinste afstand (in dit geval vrijwel altijd gelijk aan 0), dan wordt hieruit random een donor gekozen. Een belangrijke reden om binnen deze groep random te kiezen is om te voorkomen dat elke keer hetzelfde record gekozen wordt, waardoor je vertekening kunt krijgen. Een andere methode zou zijn om meer variabelen te gebruiken bij het berekenen van de Gower-afstand en zo een meer gelijkende donor aan te wijzen, maar hierdoor wordt de selectie van de donor niet altijd beter: men zal eerder gelijkheid krijgen op de “extra variabelen” en die worden net zo zwaar gewogen als de “belangrijke variabelen” die nu gebruikt worden. Aangezien in de LBZ geen extra variabelen beschikbaar waren die belangrijk zouden kunnen zijn voor het beter voorspellen van de diagnose, is gekozen voor de zes bovengenoemde variabelen.

2.5 Uitzonderingen op de standaard imputatiemethode

Voor een aantal te imputeren opnamen maken we uitzonderingen met betrekking tot de methode van de imputatie. Meer specifiek gaat het dan om:

1. Ontvangers met geslacht ‘onbekend’
2. Ontvangers met een leeftijd van 0 jaar (nuljarigen)
3. Ontvangers met een opname in kankerkliniek
4. Ontvangers die buiten Nederland wonen

Ontvangers met geslacht ‘onbekend’ (1) kunnen niet geïmputeerd worden volgens de standaard (basis)imputatie. De startgroepen (op basis van de variabelen zorgtype, specialisme en geslacht) kunnen niet gebruikt worden, omdat geslacht onbekend is. Daarom voeren we voor de ontvangers met onbekend geslacht een aparte imputatie uit, waarbij alleen specialisme en zorgtype overeen moeten komen tussen donor en ontvanger. De Gower-afstand functie wordt verder op dezelfde manier gebruikt als bij de basisimputatie en de waarde voor geslacht wordt van de donor overgenomen.

Bij ontvangers met een leeftijd van 0 jaar (2) heeft het weinig zin om leeftijd in jaren toe te voegen aan de Gower-afstand functie. Verder hebben opgenomen nuljarigen (met name pasgeborenen) een andere diagnoseverdeling dan oudere kinderen, waardoor een aparte imputatie wenselijk is. Om de startgroepen niet te klein te maken hebben we besloten dat voor deze imputatie alleen specialisme en zorgtype overeen moeten komen tussen ontvanger en

donor. Geslacht voegen we hier toe aan de berekening van de Gower-afstand en leeftijd in jaren vervangen we door leeftijd in maanden (deze kan variëren tussen 0 en 11 maanden).

Wanneer een opname plaatsvindt in een kankerkliniek (3), dan gaat het meestal om zeer specialistische zorg (behandeling kanker). Daarom worden ontvangers met een opname in de kankerkliniek niet geïmputeerd volgens de standaard (basis)imputatie. Donoren worden gekozen uit complete opnamen met oncologische diagnosen (ICD-hoofdstukken C en D) van academische ziekenhuizen. Nuljarigen met een opname in de kankerkliniek worden meegenomen in deze imputatie (en dus niet bij de aparte nuljarigen imputatie); door het gebruik van donoren met oncologische diagnosen vindt ook voor deze groep de gewenste differentiatie plaats. De Gower-afstand functie wordt berekend op basis van leeftijd, BRP migratieachtergrond, overlijden in ziekenhuis, opnameduur en urgentie van de opname.

Ten slotte worden ook de ontvangers die buiten Nederland wonen (4) apart geïmputeerd. Deze keuze is gemaakt omdat opnamen van buitenlanders vaak andersoortige opnamen betreft, bijvoorbeeld meer acute opnamen voor bepaalde diagnosegroepen. Voor niet-acute zorg wachten zij waarschijnlijk liever tot zij terug zijn in hun land van vestiging. De Gower-afstand functie wordt voor deze groep wel op dezelfde manier gebruikt als bij de basisimputatie. Incomplete opnamen van nuljarigen die buiten Nederland wonen worden bij de nuljarigen geïmputeerd; incomplete opnamen van mensen die buiten Nederland wonen bij de kankerkliniek worden bij de kankerkliniek geïmputeerd.

In Tabel 1 staan de vijf verschillende imputatiemethoden onder elkaar: per imputatiemethode wordt beschreven wat voor soort incomplete opnamen (ontvangers) en complete opnamen (donoren) voor kunnen komen.

Tabel 1. De vijf imputatiemethoden met informatie over welke eigenschappen de ontvangers en donoren hebben.

Type imputatie	**INCOMPLETE OPNAMES (ontvangers)**					**COMPLETE OPNAMES (donoren)**			
	geslacht is bekend	*nuljarigen komen voor*	*soort ziekenhuis waar opname plaats vindt*	*ontvanger woont buiten NL*		*geslacht is bekend*	*nuljarigen kunnen donor zijn*	*soort ziekenhuis waar opname plaats vindt*	*donor woont buiten NL*
basisimputatie	v	x	niet kankerkliniek	x		v	x	*	x
geslacht onbekend	x	v	*	*		v	v	*	x
*nuljarigen **	v	v	niet kankerkliniek	*		v	v	niet kankerkliniek	x
opname kankerkliniek	v	v	kankerkliniek	*		v	v	acad. ziekenhuis	x
woont buiten NL	v	x	niet kankerkliniek	v		v	x	*	v

** bij deze opnamen vindt er geen selectie plaats op dit criterium; alle opties kunnen dus voorkomen.*
*** bij nuljarigen wordt geslacht meegenomen bij de berekening van de Gower-afstand; dan hoeft geslacht dus niet altijd gelijk te zijn tussen donor en ontvanger.*

3. Kwaliteitsschatting

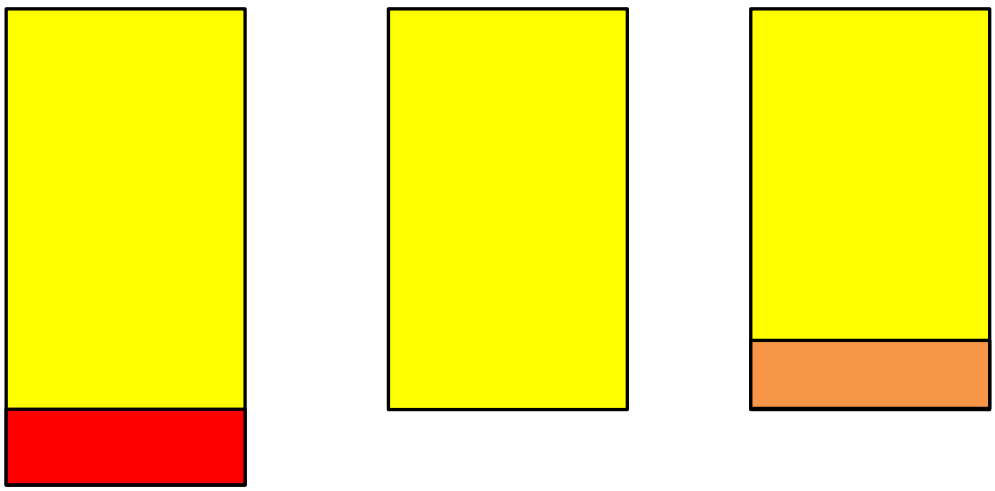
In deze paragraaf beschrijven we hoe we de kwaliteit van in paragraaf 2 beschreven nearest-neighbour imputatiemethode met de Gower-afstand hebben geschat en wat de resultaten hiervan zijn.

3.1 De aanpak

Een validatie is uitgevoerd op twee groepen: de basisimputatie en de imputatie van ontvangers met een leeftijd van 0 jaar (nuljarigen). Om de validatie uit te voeren is gebruik gemaakt van LBZ-gegevens van 2015 die volledig waargenomen zijn (complete opnamen). Voor deze groep zijn de percentages incomplete opnamen (missings) uit 2013 (per imputatiegroep) genomen: dit betreft 31,5% van de klinische opnamen en 15,8% van de dagopnamen. Vervolgens zijn random complete records van 2015 als incompleet (missing) aangewezen (bij deze records zijn de diagnosen leeg gemaakt): dit noemen we “synthetisch gaten aanbrengen”. Verder zijn voor 20% van deze records de BRP migratieachtergrond en de urgentie van de opname leeggemaakt (omdat dit ook het geval was in de LBZ-gegevens van 2013). Door de synthetische gaten te imputeren en de geïmputeerde waarden te vergelijken met de werkelijke diagnosen kan de kwaliteit van de imputatie gemeten worden.

We illustreren de nieuwe situatie in Figuur 1. In ons voorbeeld worden in 23% van de records synthetisch gaten aangebracht (oranje gemarkeerde records rechts in Figuur 1). In de overige records zijn geen synthetische gaten aangebracht (de gele records rechts in Figuur 1); dit zijn de complete records in ons voorbeeld.

Figuur 1. Aanpak “synthetisch gaten aanbrengen”.
Links: Originele data: complete (geel) en incomplete records (rood);
Midden: Alleen complete data (geel) wordt gebruikt voor de kwaliteitsschatting;
Rechts: Complete records (geel) (zoals origineel) en records met synthetische gaten (oranje gemarkeerd).



De verhouding tussen het aantal oranje records en het aantal gele records in de synthetische data van 2015 (rechts in Figuur 1) is gelijk aan de verhouding tussen het aantal rode records en het aantal gele records in de data van 2013. De rechterkant van Figuur 1 is in feite een soort miniatuurversie van de linkerkant van Figuur 1 (behalve dat voor de fractie incomplete opnamen het percentage van 2013 is aangehouden), met het grote verschil dat we de echte waardes voor het records met de synthetische gaten (het oranje gemarkeerde gedeelte links in Figuur 1) weten.

Na constructie van de synthetische data imputeren we de records met synthetische gaten met behulp van de imputatieprocedure uit paragraaf 2.

LET OP: De records die we als te imputeren records aanwijzen (de synthetische gaten) worden volledig random getrokken. Dit is ook terug te zien in de resultaten. In de echte LBZ-gegevens kunnen er echter redenen zijn dat bepaalde groepen selectief onvolledig zijn, bijvoorbeeld een bepaald specialisme in een ziekenhuis dat onvolledig heeft geregistreerd. Deze mogelijke selectiviteit wordt niet meegenomen in deze validatie.

3.2 Vertekening ten gevolge van imputatie

3.2.1 Populatieschatting van geïmputeerden

Omdat de ICD10 hoofddiagnose de belangrijkste te imputeren (categorische) variabele is, vergelijken we de resultaten eerst voor de hoofddiagnose. Omdat er zeer veel (circa 9.000) verschillende ICD10-codes zijn, bekijken we de resultaten op ICD-hoofdstukniveau (21 categorieën).

Om de mogelijke vertekening ten gevolge van de imputatieprocedure te schatten nemen we aan dat de ontvangers sterk op de donoren lijken. In het bijzonder nemen we aan dat de synthetische gaten binnen een groep random over de records in de groep zijn verdeeld. Het totaal aantal opnamen per ICD-hoofdstuk (voordat we synthetische gaten schieten) voor het oranje gemarkeerde deel in *Figuur 1 Rechts* noemen we het “Origineel aantal”. Het totaal aantal opnamen per ICD-hoofdstuk na imputatie noemen we de “Populatieschatting geïmputeerden” (dit gaat eveneens over de oranje gemarkeerde opnamen in *Figuur 1 Rechts*). Hoe kleiner het verschil tussen beide aantallen is, des te kleiner is de mogelijke vertekening ten gevolge van de imputatieprocedure (en dus hoe beter de kwaliteit van de imputatie).

We bekijken de kwaliteit van de imputatie dus alleen binnen de groep geïmputeerden. Aangezien de totale populatie opnamen inclusief de niet geïmputeerde groep opnamen (met geregistreerde diagnosen) veel groter is (het rode én gele gedeelte in *Figuur 1 Rechts)* zal de relatieve vertekening bij de overall populatieschattingen veel minder zijn dan bij de populatieschattingen van de geïmputeerde groep.

3.2.2 Resultaten

De resultaten voor de ICD-hoofdstukken staan in Tabel 2a (dagopnamen) en Tabel 2b (klinische opnamen).

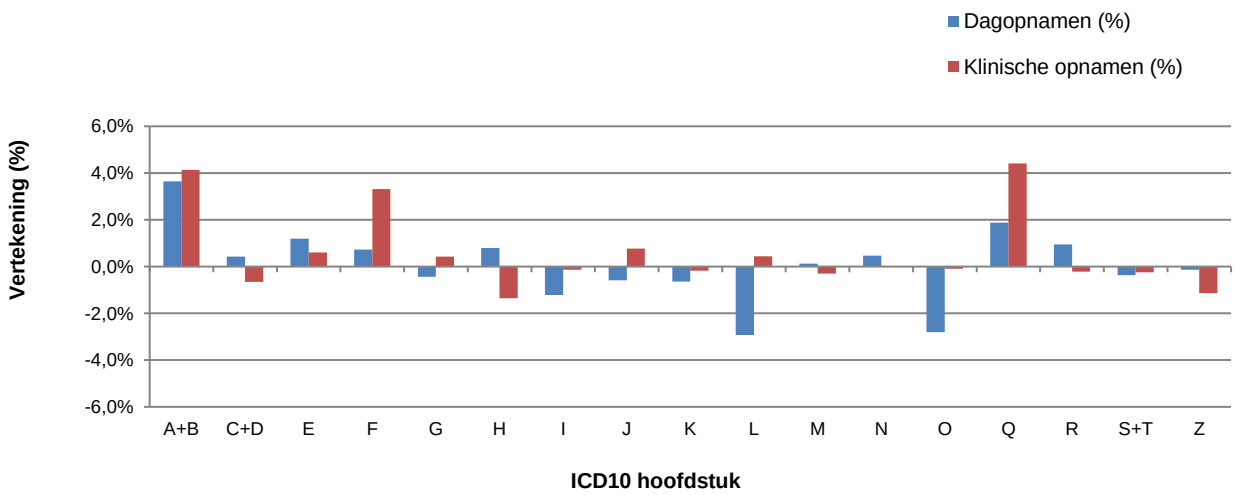
In deze tabellen staan per ICD-hoofdstuk:

- het originele aantal (zoals beschreven in 3.2.1)
- de populatieschatting geïmputeerden (zoals beschreven in 3.2.1)
- de vertekening voor de schatting (het verschil tussen het werkelijke en het aantal opnamen per ICD-hoofdstuk na de imputatie)
- de vertekening als percentage (%)

Omdat ICD-hoofdstuk P ("Bepaalde aandoeningen die hun oorsprong hebben in perinatale periode") in de synthetische data erg weinig opnamen heeft (19 dagopnamen en 27 klinische opnamen), hebben we deze niet meegenomen in de berekeningen van de vertekening.

De gemiddelde vertekening (som van de vertekeningen gedeeld door het totaal aantal originele opnamen) is voor de dagopnamen 0,61%; voor de klinische opnamen is dit 0.47%. Dit is nagenoeg verwaarloosbaar. De grootste vertekening (4,41%) treedt op bij de klinische opnamen met een diagnose in ICD-hoofdstuk Q (congenitale afwijkingen, misvormingen en chromosoomafwijkingen) (zie Figuur 2). De daaropvolgende grootste vertekening treedt op bij ICD-hoofdstukken A+B (bepaalde infectieziekten en parasitaire aandoeningen): dit is zowel bij de klinische opnamen (4,14%) als bij de dagopnamen (3,64%) het geval. Zelfs de grootste vertekeningen zijn dus nog vrij klein.

Figuur 2. Procentuele vertekening per ICD-hoofdstuk voor de geïmputeerde dagopnamen (blauw) en klinische opnamen (rood).



Tabel 2a. Mogelijke vertekening per ICD-hoofdstuk – geïmputeerde dagopnamen.

ICD10_grp	ICD10 naam	Origineel aantal	Populatie-schatting geïmputeerden	Vertekening (aantal opnamen)	Vertekening (% van origineel aantal)
A + B	Bepaalde infectieziekten en parasitaire aandoeningen (A00-B99)	1.099	1.059	40	3,64%
C + D	Nieuwvormingen (C00-D48) + Ziekten van bloed en bloedvormende organen en bepaalde aandoeningen van immuunsysteem (D50-D89)	86.895	86.528	367	0,42%
E	Endocriene ziekten en voedings- en stofwisselingsstoornissen (E00-E90)	3.116	3.079	37	1,19%
F	Psychische stoornissen en gedragsstoornissen (F00-F99)	4.232	4.201	31	0,73%
G	Ziekten van zenuwstelsel (G00-G99)	13.044	13.101	-57	-0,44%
H	Ziekten van oog en adnexen (H00-H59) + Ziekten van oor en processus mastoideus (H60-H95)	32.626	32.365	261	0,80%
I	Ziekten van hart en vaatstelsel (I00-I99)	19.584	19.822	-238	-1,22%
J	Ziekten van ademhalingsstelsel (J00-J99)	14.968	15.057	-89	-0,59%
K	Ziekten van spijsverteringsstelsel (K00-K93)	41.273	41.537	-264	-0,64%
L	Ziekten van huid en subcutis (L00-L99)	5.667	5.833	-166	-2,93%
M	Ziekten van botspierstelsel en bindweefsel (M00-M99)	37.970	37.923	47	0,12%
N	Ziekten van urogenitaal stelsel (N00-N99)	12.612	12.553	59	0,47%
O	Zwangerschap, bevalling en kraambed (O00-O99)	5.609	5.766	-157	-2,80%
Q	Congenitale afwijkingen, misvormingen en chromosoomafwijkingen (Q00-Q99)	2.452	2.406	46	1,88%
R	Symptomen, afwijkende klinische bevindingen en laboratoriumuitslagen, niet elders geclassificeerd (R00-R99)	16.343	16.187	156	0,95%
S + T	Letsel, vergiftiging en bepaalde andere gevolgen van uitwendige oorzaken (S00-T98)	6.746	6.771	-25	-0,37%
Z	Factoren die de gezondheidstoestand beïnvloeden en contacten met gezondheidszorg (Z00-Z99)	39.001	39.056	-55	-0,14%

Tabel 2b. Mogelijke vertekening per ICD-hoofdstuk – geïmputeerde klinische opnamen.

ICD10_grp	ICD10 naam	Origineel aantal	Populatie-schatting geïmputeerden	Vertekening (aantal opnamen)	Vertekening (% van origineel aantal)
A + B	Bepaalde infectieziekten en parasitaire aandoeningen (A00-B99)	4.641	4.449	192	4,14%
C + D	Nieuwvormingen (C00-D48) + Ziekten van bloed en bloedvormende organen en bepaalde aandoeningen van immuunsysteem (D50-D89)	28.918	29.106	-188	-0,65%
E	Endocriene ziekten en voedings- en stofwisselingsstoornissen (E00-E90)	6.063	6.026	37	0,61%
F	Psychische stoornissen en gedragsstoornissen (F00-F99)	2.177	2.105	72	3,31%
G	Ziekten van zenuwstelsel (G00-G99)	8.611	8.575	36	0,42%
H	Ziekten van oog en adnexen (H00-H59) + Ziekten van oor en processus mastoideus (H60-H95)	2.074	2.102	-28	-1,35%
I	Ziekten van hart en vaatstelsel (I00-I99)	36.604	36.655	-51	-0,14%
J	Ziekten van ademhalingsstelsel (J00-J99)	20.847	20.687	160	0,77%
K	Ziekten van spijsverteringsstelsel (K00-K93)	22.515	22.555	-40	-0,18%
L	Ziekten van huid en subcutis (L00-L99)	2.257	2.247	10	0,44%
M	Ziekten van botspierstelsel en bindweefsel (M00-M99)	19.796	19.856	-60	-0,30%
N	Ziekten van urogenitaal stelsel (N00-N99)	14.463	14.462	1	0,01%
O	Zwangerschap, bevalling en kraambed (O00-O99)	19.480	19.497	-17	-0,09%
Q	Congenitale afwijkingen, misvormingen en chromosoomafwijkingen (Q00-Q99)	907	867	40	4,41%
R	Symptomen, afwijkende klinische bevindingen en laboratoriumuitslagen, niet elders geclassificeerd (R00-R99)	17.262	17.298	-36	-0,21%
S + T	Letsel, vergiftiging en bepaalde andere gevolgen van uitwendige oorzaken (S00-T98)	23.542	23.599	-57	-0,24%
Z	Factoren die de gezondheidstoestand beïnvloeden en contacten met gezondheidszorg (Z00-Z99)	7.460	7.545	-85	-1,14%

Vervolgens zijn kruistabellen gemaakt om te kijken of op individueel record niveau de imputatie eveneens zo goed gaat. De resultaten van deze kruistabellen staan in Tabel 3a (dagopnamen) en 3b (klinische opnamen).

In de kruistabellen kunnen we zien dat de resultaten op individueel niveau veel minder goed zijn. Veel opnamen hebben een geïmputeerde diagnose die niet gelijk is aan de originele diagnose. Maar de hoogste frequenties van de ICD10 code-combinaties (origineel en geïmputeerd) zitten wel op de diagonaal (wat betekent dat de geïmputeerde diagnose in hetzelfde ICD-hoofdstuk valt) en de gegevens zijn zeer symmetrisch ten opzichte van de diagonaal. Dit is een indicatie dat er geen systematische vertekening optreedt door de imputatie; ofwel er zijn geen onwenselijke verschuivingen naar bepaalde diagnosegroepen.

Een voorbeeld: er zijn 1.414 klinische opnamen die origineel een ICD-hoofdstuk K-code hebben en die als een ICD-hoofdstuk I-code worden geïmputeerd. Maar er zijn dus ook 1.380 klinische opnamen die origineel een ICD-hoofdstuk I-code hebben en die als een ICD-hoofdstuk K-code worden geïmputeerd. Ook op een dieper niveau is deze symmetrie zichtbaar. Deze symmetrie is te verklaren uit het feit dat per imputatiegroep (specialisme, zorgtype en geslacht) de verdeling van de donoren wordt toegepast op de incomplete opnamen, waarbij we met behulp van de Gower-afstand zoeken naar de best bijpassende donor. Het gevolg hiervan is dat de verschuivingen van ICD-hoofdstuk X naar ICD-hoofdstuk Y en omgekeerd ongeveer gelijk zullen zijn.

Tabel 3a. Kruistabel van aantallen per ICD-hoofdstuk – origineel versus geïmputeerde dagopnamen.

	geïmputeerde ICD- hoofdstuk																			
		A+B	C+D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S+T	Z	Totaal
originele ICD-hoofdstuk	A+B	59	211	8	5	158	22	25	21	161	25	123	56	34	0	5	62	21	103	1.099
	C+D	188	52.207	1.100	59	484	248	1.122	1.436	9.536	850	1.775	2.113	223	1	262	3.356	807	11.128	86.895
	E	10	1.203	439	13	134	129	45	33	338	19	129	31	10	0	17	159	38	369	3.116
	F	4	42	14	3.710	96	7	9	4	33	2	44	23	0	1	6	125	11	101	4.232
	G	121	496	119	100	6.583	455	325	778	178	83	1.747	33	2	1	52	804	167	1.000	13.044
	H	18	279	153	12	469	26.022	30	4.191	66	36	175	7	0	1	217	316	244	390	32.626
	I	16	1.085	45	12	266	26	11.505	21	1.043	207	208	120	1	0	93	2.241	260	2.435	19.584
	J	21	1.446	37	11	789	4.023	34	6.714	119	32	95	11	3	0	125	846	160	502	14.968
	K	176	9.269	368	24	197	58	1.076	91	19.286	570	713	237	5	1	117	2.448	1.152	5.485	41.273
	L	14	790	15	3	108	33	217	28	562	2.168	701	86	6	0	63	78	294	501	5.667
	M	146	1.776	119	42	1.812	204	190	63	675	742	26.432	123	4	0	156	829	1.486	3.171	37.970
	N	50	2.105	45	18	34	7	116	2	258	72	95	5.721	1.389	2	355	463	173	1.707	12.612
	O	28	219	11	0	5	0	2	0	6	14	2	1.278	3.317	2	22	77	28	598	5.609
	P	0	1	0	0	2	0	0	0	0	0	0	3	9	0	2	0	0	2	19
	Q	8	245	31	10	41	209	92	123	154	65	167	359	10	0	526	74	107	231	2.452
	R	79	3.365	157	87	791	334	2.157	872	2.479	82	906	463	67	0	94	2.123	147	2.140	16.343
	S+T	27	769	41	7	173	207	320	143	1.140	347	1.487	142	25	0	84	124	840	870	6.746
	Z	94	11.020	377	88	959	381	2.557	537	5.503	519	3.124	1.747	661	3	210	2.062	836	8.323	39.001

Tabel 3b. Kruistabel van aantallen per ICD-hoofdstuk – origineel versus geïmputeerde klinische opnamen.

	geïmputeerde ICD-hoofdstuk																			
		A+B	C+D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S+T	Z	Totaal
originele ICD-hoofdstuk	A+B	374	596	235	54	84	24	383	713	657	81	111	354	24	0	14	443	448	46	4.641
	C+D	573	10.841	1.008	94	737	121	1.938	2.048	3.018	215	704	3.083	325	0	86	1.529	1.693	905	28.918
	E	257	1.019	836	78	106	22	473	471	1.097	91	160	306	12	1	16	439	541	138	6.063
	F	44	105	72	904	120	5	153	105	93	9	62	73	8	0	6	167	200	51	2.177
	G	65	828	97	92	3.035	241	1.146	470	113	15	626	71	13	0	23	808	617	351	8.611
	H	23	134	25	12	221	597	156	486	36	12	57	13	1	1	20	134	108	38	2.074
	I	354	2.145	421	139	1.137	178	20.583	1.552	1.380	124	567	431	14	0	55	4.134	1.790	1.600	36.604
	J	698	1.981	497	115	506	484	1.619	10.941	747	98	242	569	16	1	56	1.505	621	151	20.847
	K	645	2.963	1.089	99	119	38	1.414	791	8.216	456	373	721	54	0	55	1.481	3.488	513	22.515
	L	71	198	91	12	18	12	128	107	467	196	106	127	9	0	14	113	486	102	2.257
	M	103	648	134	53	675	49	567	235	330	82	13.459	170	4	0	83	305	2.521	378	19.796
	N	354	3.073	360	68	55	16	491	546	728	130	177	5.715	677	0	60	885	656	472	14.463
	O	27	269	22	5	5	0	8	11	54	6	7	739	17.310	6	5	169	87	750	19.480
	P	0	3	1	0	0	0	0	2	0	1	0	0	13	0	5	2	0	0	27
	Q	8	92	21	4	39	31	53	52	44	9	80	76	7	1	241	43	64	42	907
	R	439	1.559	436	132	795	135	4.156	1.432	1.557	125	333	862	165	3	31	3.510	1.183	409	17.262
	S+T	370	1.731	558	195	580	114	1.764	560	3.533	508	2.452	690	101	0	62	1.216	8.596	512	23.542
	Z	44	921	123	49	343	35	1.623	165	485	89	340	462	744	0	35	415	500	1.087	7.460

Er is een aantal ICD-hoofdstukken met relatief lage aantallen op de diagonaal (zie Tabel 4), bijvoorbeeld bij de ICD-hoofdstukken:

- A+B (Bepaalde infectieziekten en parasitaire aandoeningen): *bij dag- en klinische opnamen*
- E (Endocriene ziekten en voedings- en stofwisselingsstoornissen): *bij dag- en klinische opnamen*
- L (Ziekten van huid en subcutis): *bij klinische opnamen*
- R (Symptomen, afwijkende klinische bevindingen en laboratoriumuitslagen, niet elders geclassificeerd): *bij dagopnamen*
- S+T (Letsel, vergiftiging en bepaalde andere gevolgen van uitwendige oorzaken): *bij dagopnamen*
- Z (Factoren die de gezondheidstoestand beïnvloeden en contacten met gezondheidszorg): *bij klinische opnamen*

Lage aantallen op de diagonaal treden op bij specialismen waar een ICD10 code relatief weinig voorkomt. De oorspronkelijke waarden worden uitgesmeerd over alle codes, en alle codes dragen bij aan de nieuwe geïmputeerde waarneming. Dit is niet te voorkomen, aangezien er voor het berekenen van de Gower-afstand geen extra variabelen beschikbaar waren die de diagnose beter zouden kunnen voorspellen.

Tabel 4. Percentage opnamen per ICD-hoofdstuk op de diagonaal (zie tabellen 3a en 3b), uitgesplitst voor dagopnamen en klinische opnamen.

ICD10_grp	ICD10 naam	Opnamen op diagonaal (%)	
		Dagopnamen	*Klinische opnamen*
A + B	Bepaalde infectieziekten en parasitaire aandoeningen (A00-B99)	5,4	8,1
C + D	Nieuwvormingen (C00-D48) + Ziekten van bloed en bloedvormende organen en bepaalde aandoeningen van immuunsysteem (D50-D89)	60,1	37,5
E	Endocriene ziekten en voedings- en stofwisselingsstoornissen (E00-E90)	14,1	13,8
F	Psychische stoornissen en gedragsstoornissen (F00-F99)	87,7	41,5
G	Ziekten van zenuwstelsel (G00-G99)	50,5	35,2
H	Ziekten van oog en adnexen (H00-H59) + Ziekten van oor en processus mastoideus (H60-H95)	79,8	28,8
I	Ziekten van hart en vaatstelsel (I00-I99)	58,7	56,2
J	Ziekten van ademhalingsstelsel (J00-J99)	44,9	52,5
K	Ziekten van spijsverteringsstelsel (K00-K93)	46,7	36,5
L	Ziekten van huid en subcutis (L00-L99)	38,3	8,7
M	Ziekten van botspierstelsel en bindweefsel (M00-M99)	69,6	68,0
N	Ziekten van urogenitaal stelsel (N00-N99)	45,4	39,5
O	Zwangerschap, bevalling en kraambed (O00-O99)	59,1	88,9
Q	Congenitale afwijkingen, misvormingen en chromosoomafwijkingen (Q00-Q99)	21,5	26,6
R	Symptomen, afwijkende klinische bevindingen en laboratoriumuitslagen, niet elders geclassificeerd (R00-R99)	13,0	20,3
S + T	Letsel, vergiftiging en bepaalde andere gevolgen van uitwendige oorzaken (S00-T98)	12,5	36,5
Z	Factoren die de gezondheidstoestand beïnvloeden en contacten met gezondheidszorg (Z00-Z99)	21,3	14,6

3.3 Aantal keren dat donoren worden gebruikt

Bij het gebruik van een hot-deckmethode is het over het algemeen aan te bevelen dat eenzelfde donor niet te vaak wordt gebruikt. Als een donor-record bijvoorbeeld 100 keer of vaker voor het imputeren van ontvanger-records wordt gebruikt, dan zouden de resultaten van de imputatieprocedure immers te veel richting de kenmerken van dat ene donor-record kunnen verschuiven. De overgrote meerderheid van de donoren moet bij voorkeur slechts één of twee keer voor het imputeren van ontvanger-records worden gebruikt, en de meerderheid van die donoren zou bij voorkeur slechts één keer moeten worden gebruikt.

In Tabel 5 staat hoe vaak donoren in de imputatie van 2015 worden gebruikt. De overgrote meerderheid van de donorrecords wordt dus slechts enkele keren gebruikt. Een aantal donorrecords wordt vaker gebruikt. Dit komt echter slechts een verwaarloosbaar aantal keer voor. In de kolom “cumulatief % van het aantal donorrecords” staat het cumulatieve percentage dat met de donoren is geïmputeerd. Circa 99,6% van de geïmputeerde records is dus geïmputeerd met een donor die maximaal 3 keer is gebruikt.

Tabel 5. Frequenties van aantallen keren dat donoren zijn gebruikt.

Aantal keer als donor gebruikt	Frequentie	% van het aantal donorrecords	Cumulatief % van het aantal donorrecords
1	160.457	88,84	88,84
2	17.539	9,71	98,55
3	1.967	1,09	99,64
4	328	0,18	99,82
5	120	0,07	99,88
6	54	0,03	99,91
7	41	0,02	99,94
8	22	0,01	99,95
9	12	0,01	99,96
10	11	0,01	99,96
11-24	55	0,03	99,99
25-49	11	0,01	100,00
50-99	4	0,00	100,00
>100	0	0,00	100,00

3.4 Aantal keren dat Gower-afstand > 0 is

Van alle geïmputeerd opnamen (205.768) waren er in 2015 slechts 3.822 waarbij de Gower-afstand (afstand tussen donor en ontvanger) groter was dan 0. Dit betekent dat in die gevallen minimaal één van de variabelen die gebruikt worden om de beste donor te selecteren, verschilde tussen donor en ontvanger. In Tabel 6 staat hoe vaak de variabelen verschilden tussen donor en ontvanger; de variabelen urgentie, overlijden in ziekenhuis, soort ziekenhuis en BRP migratieachtergrond komen bijna altijd overeen. De variabelen leeftijd en opnameduur verschillen vaker tussen donor en ontvanger. Waarschijnlijk komt dit doordat verschillen in numerieke variabelen een minder grote impact hebben op de Gower-afstand dan verschillen in nominale variabelen door de grotere range van mogelijkheden binnen de variabele (zie

voorbeeld in paragraaf 2.4). Verder is het inhoudelijk ook veel minder van betekenis voor de kans op een bepaald soort opname als donor en ontvanger iets verschillen in leeftijd en opnameduur. Voor de statistieken is het belangrijk dat de nominale variabelen zo veel mogelijk overeenkomen (urgentie, overlijden in ziekenhuis, soort ziekenhuis en BRP migratieachtergrond); en deze komen ook bijna altijd overeen.

Tabel 6. Hoe vaak verschillen de variabelen die gebruikt worden om de Gower-afstand te berekenen tussen donor en ontvanger.
(n=3.822 opnamen: percentages van het totaal aantal opnamen met Gower-afstand > 0)
De percentages in deze tabel tellen niet op tot 100%, omdat sommige donoren en ontvangers op meer dan 1 variabele verschillen.

Variabele	% opnamen met verschillen in variabele
Urgentie	0,2
Overlijden in ziekenhuis	0,4
Soort ziekenhuis	1,6
Opnameduur	22,4
BRP migratieachtergrond	0,1
Leeftijd	78,7

4. Conclusie

Op basis van de kwaliteitsschatting in deze notitie kunnen we concluderen dat de huidige imputatieprocedure (een combinatie van random hot deck en nearest-neighbour imputatie) goede resultaten geeft op populatieniveau. De mogelijke vertekening van de populatieschattingen ten gevolge van de imputatieprocedure is klein: frequentieverdelingen van de diagnosen blijven goed behouden, de meerderheid van de donoren wordt slechts enkele keren gebruikt, en de belangrijkste (nominale) variabelen komen meestal overeen tussen donor en ontvanger. Deze resultaten zijn wel gebaseerd op random gekozen synthetische gaten die gemaakt zijn in complete LBZ-records. In werkelijkheid kan bij incomplete LBZ-gegevens ook selectieve uitval voorkomen, bijvoorbeeld bij één specifiek specialisme dat onvolledig heeft geregistreerd. Deze mogelijke selectiviteit is niet meegenomen in de kwaliteitsschatting; dit zou de vertekening groter kunnen maken.

Om de vertekening ten gevolge van de imputatieprocedure te bepalen zijn de geïmputeerde diagnose van de synthetische incomplete records vergeleken met de oorspronkelijke diagnose van deze records. Deze vergelijking is alleen niet gemaakt op basis van de individuele ICD10-codes (omdat er teveel verschillende soorten codes zijn), maar op basis van de 21 ICD-hoofdstukken. In de kwaliteitsschatting was ons criterium voor "*goed*" een imputatie waarbij de ontvanger een diagnose kreeg uit hetzelfde ICD-hoofdstuk. Maar er is natuurlijk binnen een hoofdgroep een verscheidenheid aan diagnosen. Het zou daarom meer informatief kunnen zijn om eenzelfde analyse te doen op een lager niveau (ofwel meer groepen), bijvoorbeeld op basis van de "*International Shortlist of Hospital Morbidity Tabulation*" (ISHMT). Deze diagnose

indeling heeft 130 verschillende diagnosegroepen, wat een preciezer beeld kan geven van de eventuele vertekening. Ter aanvulling op de gepresenteerde analyse hebben we aan de hand van deze meer uitgebreide ISHMT indeling vastgesteld dat de vertekening nog altijd nagenoeg verwaarloosbaar is (de gemiddelde vertekening voor de dagopnamen over 130 diagnosen groepen is 1,94%; voor de klinische opnamen is dit 1,89%). Omdat deze analyse niet tot nieuwe inzichten heeft geleid, hebben we besloten de uitkomsten voor de verschillende diagnosen groepen niet aan het rapport toe te voegen.

Bij de kwaliteitsschatting van de ontwikkelde imputatieprocedure is enkel gekeken naar een selectie van complete records waar we synthetische gaten in hebben geschoten, en die we vervolgens hebben geïmputeerd. In het volledige LBZ-databestand zijn de meeste records echter compleet en hoeft maar een beperkt percentage van de records te worden geïmputeerd. Bij statistieken waarbij de uitkomsten van alle records in ogenschouw worden genomen, is de invloed van de imputatieprocedure op de uitkomsten dan ook een stuk minder groot.

De imputatieprocedure is ontwikkeld met behulp van LBZ-gegevens van 2013 en getest met LBZ-gegevens van 2015. Uit de nieuwste LBZ-leveringen van 2016 blijkt dat steeds meer klinische opnamen compleet geregistreerd worden (99.9% in 2016). Imputatie voor de klinische opnamen is vanaf dit jaar dus eigenlijk niet meer nodig. Voor recente jaren wordt de imputatie dus voornamelijk uitgevoerd voor de dagopnamen (in 2016 was 17,3% van de dagopnamen incompleet geregistreerd).

Gebruik van geïmputeerde data voor longitudinaal onderzoek op individueel niveau raden wij af. Zoals beschreven in paragraaf 3 valt weliswaar een substantieel deel van de geïmputeerde diagnosen in hetzelfde ICD-hoofdstuk als de originele diagnosen, maar het komt ook vaak voor dat een donor met een diagnose uit een andere ICD-hoofdgroep wordt gekozen. Verder houden we bij de imputatie geen rekening met meerdere opnamen bij dezelfde persoon, waardoor de geïmputeerde diagnosen vreemde ziektepatronen zouden kunnen geven op individueel niveau. Daarom raden we aan de geïmputeerde gegevens alleen voor populatieschattingen te gebruiken en niet bij (longitudinale) analyses op individueel niveau.

Referenties

[1] R Core Team (2016), R: A Language and Environment for Statistical Computing, *https://www.R-project.org*

[2] Pavoine, S., J. Vallet, A.-B. Dufour, S. Gachet and H. Daniel (2009), On the Challenge of Treating Various Types of Variables: Application for Improving the Measurement of Functional Diversity. *Oikos 118*, pp. 391–402

[3] *http://www.clustan.talktalk.net/gower_similarity.html*

[4] *https://stat.ethz.ch/education/semesters/ss2012/ams/slides/v4.2.pdf*

[5] D'Orazio M. (2016), StatMatch: Statistical Matching. R package version 1.2.4. *http://CRAN.R-project.org/package=StatMatch*

Verklaring van tekens

Niets (blanco)	Een cijfer kan op logische gronden niet voorkomen
.	Het cijfer is onbekend, onvoldoende betrouwbaar of geheim
*	Voorlopige cijfers
**	Nader voorlopige cijfers
2017–2018	2017 tot en met 2018
2017/2018	Het gemiddelde over de jaren 2017 tot en met 2018
2017/'18	Oogstjaar, boekjaar, schooljaar enz., beginnend in 2017 en eindigend in 2018
2015/'16–2017/'18	Oogstjaar, boekjaar, enz., 2015/'16 tot en met 2017/'18

In geval van afronding kan het voorkomen dat het weergegeven totaal niet overeenstemt met de som van de getallen.

Colofon

Uitgever
Centraal Bureau voor de Statistiek
Henri Faasdreef 312, 2492 JP Den Haag
www.cbs.nl

Prepress
Centraal Bureau voor de Statistiek

Ontwerp
Edenspiekermann

Inlichtingen
Tel. 088 570 70 70
Via contactformulier: www.cbs.nl/infoservice