Handreiking anonimiseren

Versie 27 mei 2019

Auteur: Werkgroep anonimisering Community van Data Experts Zorggegevens Informatieberaad Zorg

Leden werkgroep: Maartje Harbers (RIVM), Dick Johan van der Harst (Vektis), Paul Merkx (Vektis), Dick Odijk (NZa), Eric Schulte Nordholt (CBS), Lany Slobbe† (RIVM), Robert Verheij (Nivel), Laurens Zwakhals (RIVM).

Inhoudsopgave

Inleiding

Aanleiding voor deze handreiking

De gezondheidszorg is één van de grootste economische sectoren van ons land, waar iedere Nederlander mee te maken heeft. De behoefte aan informatie over de sector is dan ook groot, zowel bij publiek als bij professionals. Transparante datavoorziening heeft dan ook veel prioriteit. Daarbij kunnen drie informatiecircuits onderscheiden worden. Ten eerste is er een interne cyclus tussen zorgaanbieders, betalers, toezichthouders en andere stakeholders. Alleen door uitwisseling van kennis en informatie kan de kwaliteit van de geleverde zorg op peil blijven en kunnen uitkomsten van behandelingen gebruikt worden om de inzet van die behandelingen te verbeteren. Ook kunnen indicatoren voor systeem-verantwoording berekend worden.

Daarnaast is er een circuit van externe verantwoording naar beleid en publiek. Ten aanzien van beleid resulteert dit vooral in regelmatige publicatie van indicatoren, ten aanzien van de burger is er een beweging om iedere burger het recht te geven de over hem opgeslagen data te kunnen inzien.

In een samenleving die steeds meer data-gedreven functioneert, komt daar nog een derde circuit bij: publiek beschikbare open data, die een rol kunnen spelen bij de ontwikkeling van innovatieve informatieproducten voor de zorg, zoals apps, wearables etc. Het creëren van economische meerwaarde is een belangrijke reden voor de publicatie van open data bij de overheid (zie Begrippenlijst pagina 24 voor uitleg over open data.) Volgens de 'Wet hergebruik van overheidsinformatie' (Who) uit 2016 moeten overheidsorganisaties data die zijn geproduceerd voor publieke taken op verzoek van burgers of bedrijven beschikbaar stellen voor hergebruik door derden[1]. Als overheidsinformatie op eigen initiatief ter beschikking wordt gesteld als open data is dit ook hergebruik van overheidsinformatie. De Who is van toepassing op met een publieke taak belaste instellingen. Daaronder vallen diverse organisaties die werkzaam zijn op het terrein van de gezondheidszorg, zoals TNO, academische ziekenhuizen en het College voor Zorgverzekeringen[2]. Informatie van onderzoeksinstellingen is echter uitgesloten van de Who. In dit verband is ook het juridisch kader van de Wet op het Centraal Bureau voor de Statistiek van belang[3]. In de CBS-wet staat dat het CBS zo veel als mogelijk gebruik moet maken van gegevens uit bestaande registers die in verband met de uitvoering van een wettelijke taak worden bijgehouden bij overheden en buiten de overheid. Alleen als deze informatie niet voldoet, mag het CBS bedrijven verplichten om hun gegevens te verstrekken aan het CBS.

Alle drie 'circuits' maken deel uit van een 'lerend zorgsysteem'[4]. In een lerend zorgsysteem worden gegevens gegenereerd als onderdeel van het zorgproces. Die gegevens worden weer hergebruikt voor het vergroten van kennis over dat zorgproces. Vervolgens kan die kennis weer worden gebruikt voor veranderingen in dat zorgproces[5] [6]. De Kamerbrief "Data laten werken voor gezondheid" past ook goed bij het concept van het lerend zorgsysteem en noemt gegevensuitwisseling een randvoorwaarde voor goede zorg[7]. Een van de thema's in de brief is datasolidariteit: het delen en beschikbaar stellen van data ten behoeve van het grotere geheel, mits de doelen van de data-analyses goed gedefinieerd zijn en duidelijk meerwaarde opleveren voor de gezondheid van anderen. De beweging naar transparantie en open data kan echter in botsing komen met een andere kernwaarde uit ons zorgsysteem: het recht van

een burger en patiënt op privacy en vertrouwelijke omgang met zijn data. Gezondheidsdata van personen zijn (bijzondere) persoonsgegevens en mogen niet zomaar als open data worden gepubliceerd. De eisen aan de omgang met persoonsgegevens zijn nog versterkt door de inwerkingtreding van de Algemene Verordening Gegevensbescherming (AVG) in mei 2018. Het is dus wenselijk waar mogelijk data te delen, maar tegelijkertijd moet worden voorkomen dat er persoonsgegevens uit datasets te herleiden zijn. Dit leidt in de praktijk soms tot verlamming. Data worden niet gepubliceerd of beschikbaar gesteld aan derden vanwege privacy-onthullingsrisico's. Een neveneffect daarvan is dat die data dan ook niet beschikbaar zijn voor verbetering van de zorg (circuit 1), verantwoording (circuit 2), en/of voor innovatieve productontwikkeling (circuit 3).

Daarom moet goed worden nagedacht hoe in het zorg- en gezondheidsdomein invulling kan worden gegeven aan het open beschikbaar stellen van data. Daarbij is het aan te bevelen om publicatie van open data niet als los staand te zien van de eerste twee data- en informatiecircuits. Pseudonimisering kan een manier zijn om de privacy te beschermen in de eerste twee informatiecircuits, maar het is geen oplossing voor het open beschikbaar stellen van data met het oog op het derde circuit, omdat volgens de AVG gepseudonimiseerde gegevens nog steeds persoonsgegevens zijn en deze mogen niet zomaar ingezien of gedeeld worden. Voor open data zijn anonimisering of aggregatie van data wel mogelijke oplossingen mits herleidbaarheid naar personen is uitgesloten.

Doel en doelgroep

Deze handreiking is bedoeld als hulpmiddel voor organisaties op het gebied van zorg en gezondheid die overwegen data open te publiceren, en daarbij hun weg proberen te zoeken tussen enerzijds transparantie en anderzijds de noodzakelijke privacybescherming van de personen op wie hun data betrekking hebben. Anonimiseren (zie Begrippenlijst pagina 24 voor uitleg) kan hiervoor een oplossing bieden. Het is echter een lastige opgave om enerzijds een dataset volledig te anonimiseren en elk risico op onthulling van persoonsgegevens uit te sluiten en anderzijds te zorgen dat de overgebleven data voldoende van waarde is voor verder onderzoek. Hét juiste antwoord bestaat hier niet, want dit verschilt van geval tot geval. Om hiervoor handvatten te geven is op verzoek van het Informatieberaad Zorg deze 'Handreiking anonimiseren' opgesteld.

Hoewel de focus op het anonimiseren van microdata ligt, is dit slechts één van de mogelijkheden om data te kunnen hergebruiken. Het anonimiseren ten behoeve van het open publiceren van data moet worden bezien in het bredere kader van de FAIR-principes. Dit zijn algemene principes voor goed databeheer en het geschikt maken van data voor hergebruik. FAIR staat voor findable (vindbaar), accessible (toegankelijk), interoperable (uitwisselbaar), re-usable (herbruikbaar)[8]. Open data zou moeten voldoen aan de FAIR-principes, maar data die voldoet aan FAIR hoeft niet per se open gepubliceerd te zijn. FAIR laat namelijk ruimte voor het aanbrengen van beperkingen op het hergebruik zodat persoonsgegevens volgens de eisen van de AVG kunnen worden beschermd en misbruik kan worden voorkomen. Een voorbeeld hiervan is het op basis van contracten beschikbaar stellen van microdata voor onderzoekdoeleinden al of niet na pseudonimisering door een Trusted Third Party (TTP). Een Trusted Third Party of 'vertrouwde derde partij' is een onafhankelijke partij die er voor zorgt dat de privacy beschermd blijft bij de uitwisseling van data tussen twee of meerdere verschillende partijen. De TTP kan bijvoorbeeld de sleutel in bewaring houden bij pseudonimisering van databestanden met

privacygevoelige informatie. Bij open data is geen sprake van beperkingen, zoals bij FAIR data dus wel het geval kan zijn. (zie Begrippenlijst pagina 24 voor uitleg over pseudonimisering.)

Reikwijdte

Deze handreiking beoogt enerzijds een denkhulp te zijn voor discussies over de inzet van anonimiseren. Anderzijds geeft het ook meer praktische handvatten voor anonimiseren. Het is niet bedoeld als alomvattende beschrijving van alle aspecten van anonimiseren in relatie tot open data. Hiervoor verwijzen we naar andere publicaties. Zo heeft de Technische Universiteit Delft in opdracht van het RIVM een 'beslisboom open data' ontwikkeld voor het juridische en inhoudelijke proces dat voorafgaat aan publicatie van open data[9]. Nadrukkelijk buiten deze handreiking valt anonimiseren van data van niet-natuurlijke rechtspersonen als bedrijven en organisaties. Het gaat met andere woorden over gegevens van individuen. De techniek van pseudonimiseren en het gebruik van BSN-nummers is onvoldoende voor het waarborgen van anonimiteit. Daarom vallen pseudonimiseren en het gebruik van BSN-nummers buiten het kader van deze handreiking. Tot slot dient opgemerkt te worden dat het bij het open publiceren van data niet gaat om complete databases, maar om specifieke onderzoeksbestanden en (sets van) statistische tabellen al of niet bij elkaar gebracht in een data repository.

Leeswijzer

Een dik handboek zal in de praktijk nauwelijks uit de kast gehaald worden. Daarom is er naar gestreefd de omvang van deze handreiking beperkt te houden en waar mogelijk naar andere literatuur te verwijzen. Wat het wel wil zijn is een checklist: hebben we aan dit aspect gedacht? Daarom volgt na een korte uiteenzetting over wat anonimiseren is, een stapsgewijs overzicht met aanbevelingen. Het meest leerzaam is de ervaring van organisaties die al ervaring hebben met het publiceren van open data op het terrein van zorg en gezondheid. Vier voorbeelden hiervan zijn in deze handreiking opgenomen: Vektis, RIVM, NZa en Nivel.

In opbouw bestaat deze handreiking uit drie delen:

1. Toelichting op anonimiseren
2. Aanbevelingen
3. Praktijkvoorbeelden

1. Toelichting op anonimiseren

Wat is anonimiseren?

Anonimiseren wordt hier opgevat als iedere bewerking die het koppelen van data aan natuurlijke personen onmogelijk maakt. Bij anonimiseren wordt geen koppelsleutel meegeleverd zodat het koppelen op persoon van informatie uit verschillende datasets niet mogelijk is. Anonimiseren is een dynamisch begrip. Werd in het verleden het verwijderen van naam en adresgegevens al als voldoende inspanning voor anonimiseren beschouwd, tegenwoordig wordt dit als onvoldoende gezien. Via geavanceerde data-analysetechnieken zijn al veel persoonskenmerken af te leiden uit andere informatie. Bijvoorbeeld het taalgebruik van personen op social media onthult vaak al leeftijd en geslacht van de

aanbieder[10]. Dergelijke data-analysetechnieken zullen naar verwachting steeds geavanceerder worden. De wetgever anticipeert op deze ontwikkeling in de eisen die aan anonimiseren van persoonsgegevens worden gesteld.

Voor deze handreiking is vooral de opvatting over anonimiseren uit de Algemene Verordening Gegevensbescherming (AVG) (EU verordening 2016/679) relevant[11]. Hierin zijn de belangrijkste regels voor de omgang met persoonsgegevens in Nederland vastgelegd. Zo mogen datasets met persoonsgegevens niet zomaar ingezien en gedeeld worden. De AVG werd op 25 mei 2018 van kracht en vervangt de Wet bescherming persoonsgegevens (Wbp). De AVG geldt voor elk gegeven betreffende een geïdentificeerde of identificeerbare natuurlijke persoon, maar niet voor anonieme gegevens. Een effectieve methode van anonimisering moet volgens de werkgroep Gegevensbescherming artikel 29 (nu European Data Protection Board) bewerkstelligen dat voor alle partijen wordt uitgesloten de mogelijkheid om[12]:

1. Een persoon te individualiseren (‘single out’);
2. Verschillende records in verband te brengen met een individu (‘linkability’);
3. Informatie over een individu af te leiden (‘inference’).

De verordening stelt in overweging 26 dat gegevens pas echt anoniem zijn (en daarmee dus niet onder de verordening vallen) als het ‘redelijkerwijs’ onmogelijk is personen te identificeren, rekening houdend met beperkingen in kosten en tijd, en met de huidige technologie en toekomstige technologische ontwikkelingen[13]. Deze overweging laat nog ruimte voor discussie over wat ‘redelijkerwijs’ is.

Verschil tussen anonimiseren en pseudonimiseren

Nadrukkelijk stelt de AVG dat gepseudonimiseerde persoonsgegevens, waarbij identificeerbare persoonsgegevens vervangen worden door een pseudoniem, moeten worden beschouwd als gegevens over identificeerbare natuurlijke personen. Ze worden dus nog steeds als persoonsgegevens gezien, omdat hierbij identificatie theoretisch altijd mogelijk blijft voor wie beschikt over de vertaaltabel tussen het pseudoniem en de onversleutelde informatie. De AVG maakt daarbij geen onderscheid tussen verschillende soorten pseudonimiseren, namelijk éénweg- (onomkeerbare) en tweewegpseudonimisatie (omkeerbare). Bij éénwegpseudonimisatie is het niet mogelijk om het pseudoniem direct terug te vertalen naar het oorspronkelijke gegeven. Niemand beschikt meer over de vertaaltabel terug waardoor het pseudoniem niet meer is terug te herleiden tot een individu. Bij tweewegpseudonimisatie is dat wel het geval. Het voordeel van pseudonimiseren is dat bestanden die op dezelfde wijze gepseudonimiseerd zijn, nog steeds te koppelen zijn. Dit vergroot de bruikbaarheid van gegevens voor onderzoek. Met de mogelijkheid tot koppelen onderscheiden zowel éénweg- als tweewegpseudonimiseren zich van anonimiseren, want bij anonimiseren is het koppelen op persoonsniveau van informatie uit verschillende bronnen niet mogelijk. Dat volgens de AVG gepseudonimiseerde gegevens nog steeds persoonsgegevens zijn, wil overigens niet zeggen dat ze in het geheel niet gebruikt of verstrekt kunnen worden aan derden. Het betekent alleen dat in dit geval de AVG van toepassing blijft, omdat de gegevens niet anoniem zijn.

2. Aanbevelingen

Wie data open beschikbaar wil stellen voor hergebruik door derden heeft, ook binnen het kader van de Wet hergebruik van overheidsinformatie, een ruime keuzevrijheid voor wat precies en in welke vorm open wordt gepubliceerd. Deze keuzes zijn van cruciaal belang voor de feitelijke mogelijkheden tot hergebruik. Tegelijkertijd moet worden voorkomen dat er persoonsgegevens uit datasets te herleiden zijn. Bij het publiceren van open data moet daarom niet alleen naar de wettelijke vereisten worden gekeken, maar moet ook een goed beeld worden gevormd van de beoogde eindgebruiker en zijn wensen. Hieronder volgt daarom een aantal aanbevelingen die kunnen helpen bij het vinden van een goede balans tussen aan de ene kant het beschikbaar en bruikbaar maken van data voor een zo groot mogelijke groep gebruikers en aan de andere kant het uitsluiten of minimaliseren van het risico op onthulling.

1. Houd rekening met generieke wensen ten aanzien van hergebruikte data

Voor een deel zullen veel hergebruikers van data gemeenschappelijke wensen hebben. Generieke wensen ten aanzien van te hergebruiken data zijn bijvoorbeeld:

- Publicatie in de vorm van computer-herbruikbare open data;
- Het gebruik van voor gebruikers herkenbare terminologie;
- Bij benamingen en classificaties zoveel mogelijk (internationaal) gestandaardiseerde classificaties gebruiken;
- Publiceren van metadata die de data zelf en hun productie beschrijven;
- Publiceren van metadata over bewerkingen uitgevoerd op de data om ze publicabel te maken, bijvoorbeeld anonimiseren en aggregeren.

Strikt genomen gelden deze eisen ook voor niet-geanonimiseerde data. Omdat het anonimiseringsproces echter invloed heeft op alle genoemde eisen, is het van belang hier vanaf het begin rekening mee te houden. Op een aantal van deze aspecten en hun relatie tot anonimiseren wordt elders in deze handreiking dieper ingegaan.

2. Houd rekening met specifieke gebruikerswensen ten aanzien van een databestand

Gebruikers zullen specifieke wensen hebben ten aanzien van het detailniveau van de door hen gewenste data. Manieren om hier een beter beeld van te krijgen zijn bijvoorbeeld focusgroepen met potentiële gebruikers, een brainstormsessie binnen de eigen organisatie over het mogelijke externe gebruiksnut van een open data publicatie en een analyse van al binnengekomen vragen naar data en de vragenstellers hierachter. De volgende vragen kunnen daarbij behulpzaam zijn:

- Welk detail wensen gebruikers van de data te zien? Denk daarbij aan:

  - Aggregatieniveau: kan uiteenlopen van records op individueel niveau tot statistische grootheden als totaalsom, gemiddelde of spreiding;
  - Temporele aspecten: jaar of maandcijfers, update-frequentie;
  - Ruimtelijke aspecten: nationale, provinciale, gemeentelijke of wijkcijfers;
  - Zorgkenmerken: type zorg, geleverd product;
  - Gebruikerskenmerken: leeftijd, geslacht, ziekte-diagnose, herkomst, opleiding, inkomen etc.
- In welke context zullen gebruikers onze cijfers tonen? Willen ze bijvoorbeeld verbanden leggen met gegevens uit andere databronnen?

Antwoorden op deze vragen geven zicht op welke data van nut kunnen zijn voor externe partijen, maar ook of een anonieme datapublicatie überhaupt zinvol is. Als sommige gebruikers bijvoorbeeld gegevens op wijkniveau zouden willen zien, waar vanuit technisch anonimisatie oogpunt gemeenteniveau het hoogst haalbare is, dan zal deze groep niet bediend kunnen worden met de geanonimiseerde open data. Naast dergelijke privacyoverwegingen zullen ook pragmatische aspecten een rol spelen, en zullen de inspanningen van de eigen organisatie voor productie van geanonimiseerde open data moeten worden afgewogen tegen de baten hiervan.

3. Bepaal het gewenste veiligheidsniveau

Met anonimiseren wordt beoogd onthulling te voorkomen. Van onthulling is sprake als uit een dataset informatie kan worden achterhaald over een herkenbaar afzonderlijk persoon of huishouden. Ook combinaties van categorieën van variabelen kunnen tot unieke of zeldzame personen leiden waarbij de kans groot is dat een afzonderlijke persoon wordt herkend en informatie over hem/haar wordt onthuld.

Bij het proces van anonimiseren zal gekozen moeten worden voor een ‘veiligheidsniveau’ ten aanzien van onthulling, bijvoorbeeld door het aantal detailkenmerken bij gegevens te beperken - denk aan het weglaten van uitsplitsingen naar leeftijd, geslacht, herkomst of regio - of door een ondergrens te stellen aan het minimaal aantal personen waar een te publiceren gegeven betrekking op moet hebben. Op de technische aspecten van voorkómen van onthulling wordt in Aanbeveling 4 van deze handreiking dieper ingegaan. Hier gaat het om het besef dat dit veiligheidsniveau niet absoluut is maar dynamisch; het kan verschillen naar de aard van het gegeven maar is ook cultureel en temporeel bepaald. Enkele voorbeelden:

- In Nederland worden gegevens over aantallen gelijkgeslachtelijke echtparen tegenwoordig gewoon gepubliceerd. In landen waar homoseksualiteit meer in de taboesfeer zit, is dit niet mogelijk.

- Daarentegen is het eigen inkomen en de betaalde belasting in Nederland een strikt privé-gegeven, waar in andere landen deze gegevens gewoon openbaar zijn[14].

- Wat kan en mag is ook niet constant in de tijd. Publiceren over gelijkgeslachtelijke huwelijken zou in Nederland nog maar enkele decennia geleden ook niet mogelijk zijn geweest. In het boekje “Je hebt wel iets te verbergen” van Maurits Martijn en Dimitri Tokmetzis staan interessante voorbeelden over hoe Google in de loop van de tijd een andere invulling is gaan geven aan de privacybescherming[15].

4. Kies de beste techniek voor bescherming tegen onthulling

Voor het anonimiseren van gegevens moet onderscheid worden gemaakt tussen microdata waarbij elk record gegevens van een eenheid (bijvoorbeeld een persoon of huishouden) bevat en statistische data (bijvoorbeeld tabellen) die geaggregeerde informatie bevatten. De eerste stap bij anonimiseren van microdata is het verwijderen van direct te herleiden kenmerken zoals namen, geboortedata en adressen. Ook moeten combinaties van specifieke kenmerken (geboortedatum/adres) die (meestal) leiden tot unieke entiteiten (personen) worden vermeden. Hoe meer kenmerken worden geleverd, hoe groter de kans op onthullingsrisico's, vooral als dit zeer specifieke en weinig voorkomende kenmerken zijn.

Bij statistische tabellen gaat het om geaggregeerde uitkomsten zoals percentages, totalen, gemiddelden en standaardafwijking. Deze tabellen bevatten dus geen aparte records meer per persoon of huishouden waardoor deze doorgaans ook geen direct tot de persoon herleidbare kenmerken als naam en adres bevatten. In die zin is dit dus zowel een vorm van anonimiseren als een alternatief voor het open publiceren van microdata (zie Aanbeveling 8). Toch zijn ook statistische data lang niet altijd direct geschikt voor publicatie, want ook deze geaggregeerde data kunnen tot onthulling leiden. Bijvoorbeeld als ze op een vrij kleine groep van toepassing zijn en deze personen elkaar kennen en onderling informatie uitwisselen. Denk aan een schoolklas: als de leraar onthult dat er één '10' is uitgedeeld, dan zal het meestal weinig moeite kosten er achter te komen wie de '10' heeft gehaald, zelfs als de betreffende leerling het anoniem zou willen houden. Er moet worden voorkomen dat door publicatie onthulling van individuele informatie kan plaatsvinden. Zeker gezien de toenemende hoeveelheid informatie die beschikbaar is op internet is dat geen gemakkelijke taak. Hieronder gaan we eerst in op technieken voor het beschermen tegen onthulling bij microdata en daarna op technieken voor het beschermen van tabellen met geaggregeerde statistische data.

Bij de keuze voor de te gebruiken methode(n) moeten het onthullingrisico en informatieverlies tegen elkaar afgewogen worden. Hoe kleiner het informatieverlies, hoe groter het onthullingrisico. En omgekeerd: hoe kleiner het onthullingsrisico hoe groter het informatieverlies. Geen enkele techniek is dus vrij van tekortkomingen. De optimale oplossing voor het anonimiseren van data moet daarom per situatie worden gekozen waarbij ook vaak meerdere methoden tegelijk gebruikt kunnen worden.

A. Technieken voor de bescherming van microdata

1. Globaal hercoderen en lokaal onderdrukken

Voor het beschermen van microdata zijn globaal hercoderen en lokaal onderdrukken twee veelgebruikte technieken[16]. Globaal hercoderen is een vorm van generalisatie waarbij eigenschappen van de betrokkenen worden veralgemeniseerd. Zo kunnen twee of meer categorieën van identificerende variabelen worden samengenomen. Een voorbeeld is dat de variabele provincie wordt vervangen door de variabele landsdeel. Ook kunnen inkomenscategorieën worden gebruikt in plaats van precieze bedragen.

Lokaal onderdrukken wil zeggen dat een zeldzame score wordt vervangen door een code onbekend. Een voorbeeld is dat het beroep 'burgemeester' wordt vervangen door beroep 'onbekend'. Met globaal hercoderen en lokaal onderdrukken kunnen dus zeldzame combinaties van identificerende variabelen minder zeldzaam gemaakt worden waardoor ze voldoende vaak in het bestand voorkomen om het risico op onthulling te minimaliseren. Bij globaal hercoderen wordt snel veel bescherming bereikt, maar gaat ook veel informatie verloren. Bij lokaal onderdrukken valt het informatieverlies relatief mee, maar wordt veel minder bescherming bereikt. Het is vaak een samenspel tussen wensen (welke categorie is van essentieel belang voor gebruikers?) en mogelijkheden (is er sprake van voldoende bescherming?) om tot een goede bescherming te komen.

Bedacht moet worden dat van een microdatabestand slechts één versie openbaar kan worden gemaakt. Als twee verschillende beveiligde versies openbaar worden gemaakt, bestaat het risico dat door combinatie van gegevens uit beide beveiligde bestanden toch individuele informatie kan worden onthuld.

2. Randomisatie

Diverse andere technieken om microdata te beschermen vallen onder de term 'randomisatie'[17]. Bij randomisatie worden de attribuutwaarden die bij individuen horen gewijzigd, waardoor de data lastiger te herleiden is tot specifieke personen. Onder randomisatie vallen ook het toevoegen van ruis aan de gegevens en het verwisselen van informatie van verschillende eenheden. Dergelijke technieken hebben als nadeel dat voor een gebruiker van een op een dergelijke manier beveiligd bestand niet duidelijk is of de informatie over een individu nog correct is. Ook is meestal niet duidelijk of er voldoende aan het bestand is veranderd voor een adequate beveiliging. Voordeel van dergelijke technieken is dat de structuur van het bestand onveranderd is (dat geldt niet bij globaal hercoderen) en er (in sommige gevallen wel aangepaste) gegevens zijn voor alle velden in het bestand (dat geldt niet bij lokaal onderdrukken). Microdatabestanden waarop randomisatie is toegepast kunnen gebruikt worden als testbestand om bijvoorbeeld scripts uit te testen. De uiteindelijke definitieve analyse zou dan vervolgens uitgevoerd kunnen worden op het originele (onbeveiligde) bestand binnen een beveiligde omgeving, al of niet via een remote access voorziening (zie Aanbeveling 8).

3. Synthetische data

Ten slotte is het mogelijk synthetische data samen te stellen en die beschikbaar te stellen in plaats van de oorspronkelijke microdata. In dergelijke synthetische data kunnen bepaalde eigenschappen behouden blijven zoals de gemiddelden en varianties van variabelen. Bij gebruik van synthetische data moet echter altijd worden bedacht dat niet alle analyses op deze data dezelfde resultaten zullen opleveren als analyses op de oorspronkelijke microdata. Daardoor zijn synthetische data minder bruikbaar voor onderzoeksdoeleinden. Synthetische data zijn echter wel bruikbaar als testbestand om bijvoorbeeld de juiste werking van software te kunnen testen.

B. Technieken voor de bescherming van statistische data

Bij de beveiliging van statistische tabellen moet onderscheid worden gemaakt tussen frequentietabellen en kwantitatieve tabellen. Voor beide typen tabellen geldt dat er vaak andere, gerelateerde tabellen zijn. Deze gerelateerde tabellen zullen simultaan beschermd moeten worden om te voorkomen dat met informatie uit de ene tabel beschermde informatie uit de andere tabel kan worden onthuld (zie Aanbeveling 6).

Voor het beveiligen van statistische tabellen zijn grofweg drie methoden beschikbaar:

- Herstructureren van de tabel: door het samenvoegen van categorieën wordt de vulling per cel vergroot;
- Onderdrukken: niet publiceren van bepaalde cellen. De celwaarde wordt dan vervangen door een kruisje (×);
- Afronden: door celwaarden af te ronden zijn de exacte celwaarden slechts binnen een bepaald interval bekend.

Voor meer details over deze methoden verwijzen we naar het rapport Statische Beveiliging[18]. Hieronder gaan we verder in op enkele aandachtspunten bij de beveiliging van frequentietabellen en kwantitatieve tabellen.

1. Frequentietabellen

Frequentietabellen zijn tabellen waar alleen maar aantallen in voorkomen. Die aantallen zijn op zich geen probleem, maar de opspanvariabelen kunnen gevoelig zijn en dan kan individuele informatie worden onthuld. Bij opspanvariabelen gaat het om de variabelen in een tabel waarnaar wordt getabelleerd. Als bijvoorbeeld in een tabel staat aangegeven dat 23 jongeren in een bepaalde wijk crimineel zijn, is dat op zich geen reden om dit gegeven niet te publiceren. Als echter in de tabel ook staat dat er in diezelfde wijk geen jongeren wonen die niet crimineel zijn, is er een publicatieprobleem. Een lezer van de tabel kan onmiddellijk concluderen dat er maar 23 jongeren in die wijk zijn en dat die allemaal crimineel zijn. In combinatie met het tweede gegeven wordt het eerste gegeven dus ook onthullend. We noemen dit ook wel groepsonthulling. Mocht de lezer een jongere in die wijk kennen dan weet hij door publicatie van deze tabel ineens dat hij een crimineel kent. Ook als er slechts 1 jongere in de betreffende wijk niet crimineel is, hebben we een publicatieprobleem. De betreffende jongere weet dan namelijk dat alle andere jongeren in zijn wijk crimineel zijn. In het geval dat twee jongeren niet crimineel zijn in die wijk is er nog steeds een probleem. Deze twee jongeren kunnen gezamenlijk concluderen dat alle andere jongeren in hun wijk crimineel zijn. Het is natuurlijk mogelijk alle scores 1 en 2 in frequentietabellen te verbieden of uit veiligheidsoverwegingen een hogere drempelwaarde te stellen (bijvoorbeeld dat elke cel minimaal 10 eenheden moet bevatten en anders niet publicabel is). Overigens maakt dit voorbeeld ook duidelijk dat er geen absolute bescherming tegen groepsonthulling mogelijk is. Als slechts één criminele jongere in de wijk zou wonen, dan kunnen de andere 22 daarachter komen door onderling informatie uit te wisselen. Hoe hoger het aantal, hoe onwaarschijnlijker dat dit

gebeurt. Daarom wordt meestal een grens van 10 adequaat geacht, maar dit kan variëren naar gelang het gewenste veiligheidsniveau (zie Aanbeveling 3).

Bedacht moet worden dat onderdrukken van individuele celwaarden geen adequate bescherming is als ook randtotalen worden gepubliceerd. In elke rij of kolom van een frequentietabel met één onderdrukte celwaarde kan gemakkelijk worden afgeleid wat die onderdrukte waarde is. Er zullen dus nog extra celwaarden moeten worden onderdrukt om tot een adequate bescherming te komen. Hoe die extra onderdrukte cellen optimaal moeten worden bepaald (dat wil zeggen met zo min mogelijk informatieverlies maar nog juist voldoende bescherming biedend) is een moeilijk wiskundig probleem dat voor tabellen van beperkte omvang met softwarepakketten als Tau-ARGUS kan worden aangepakt[19].

2. Kwantitatieve tabellen

Kwantitatieve tabellen bevatten kwantitatieve gegevens in de cellen. Hierbij kan bijvoorbeeld worden gedacht aan een tabel met gemiddelde inkomensgegevens naar geslacht en leeftijd in een bepaalde regio. Het is niet mogelijk om uit een dergelijke tabel direct individuele inkomensinformatie af te leiden omdat niet duidelijk is op hoeveel personen of huishoudens de celwaarden betrekking hebben. Zodra die aantallen echter bekend zijn, kan worden bekeken of individuele informatie kan worden onthuld. Het is duidelijk dat cellen die betrekking hebben op één eenheid niet kunnen worden gepubliceerd. Ook cellen die betrekking hebben op twee eenheden kunnen niet worden gepubliceerd. Dan zou immers de ene bijdrager (die zijn eigen inkomen kent) het inkomen van de andere bijdrager (dat hij nog niet kent) kunnen terugrekenen. Bij grotere aantallen bijdragers speelt dominantie een rol. Als er een persoon is met een heel hoog inkomen domineert die de celwaarde en kan een goede schatting van het inkomen van die persoon worden verkregen bij publicatie van de betreffende celwaarde. Er zijn regels op te stellen die aangeven dat alle individuele inkomens niet nauwkeuriger mogen worden teruggerekend dan tot een zeker percentage van de werkelijke waarde. Op basis van die regel kan dan met behulp van softwarepakketten als Tau-ARGUS worden bepaald welke celwaarden niet mogen worden gepubliceerd[20].

5. Wees je bewust van stigmatisering en voorkom foutieve interpretaties

Ook al zijn data geanonimiseerd, publicatie kan nog steeds gevolgen hebben voor individuen. Een voorbeeld is publicaties over aantallen diefstallen in een regio en andere criminaliteit: dat kan bijvoorbeeld de verkoopbaarheid van huizen negatief beïnvloeden en daarmee schadelijke gevolgen voor personen hebben. Publiceren van bijvoorbeeld zorggebruikscijfers naar persoonskenmerken als woonplaats, inkomen, leeftijd, geslacht of etnische herkomst kan tot stigmatisering van groepen leiden, met mogelijk kwalijke maatschappelijke gevolgen als vooroordelen, achterstelling of discriminatie. Publiceren dat ouderen meer zorg gebruiken dan gemiddeld is daar een nog relatief onschuldig voorbeeld van, dat publicitair toch al vaak tot felle reacties leidt.

Het risico op stigmatisering kan dus pleiten tegen publicatie van gegevens met bepaalde kenmerken. Er zijn echter ook argumenten die juist vóór publicatie pleiten:

- Eventuele negatieve gevolgen van publicatie als achterstelling of discriminatie worden al in anti-discriminatiewetgeving afgedekt;
- Het is niet wenselijk als in het kader van publieke taken verzamelde informatie niet wordt gepubliceerd. Dit staat haaks op waarden als transparantie en verantwoording;
- Niet publiceren is uiteindelijk kwalijker dan wel publiceren. Je bevordert er feitenvrije discussie mee. Vooroordelen bestaan ook zonder data en door niet te publiceren ontneem je groepen het recht vooroordelen te ontkrachten dan wel te bevestigen.

Wettelijk zijn er geen problemen met het publiceren van data naar groepskenmerken, mits personen niet herkenbaar zijn, wat in de praktijk betekent dat de statistische eenheden voldoende omvang moeten hebben. Maar de gegevensvoorbeelden geven aan dat er ook een buitenwettelijke verantwoordelijkheid ligt bij het publiceren van gegevens die niet ophoudt bij anonimiseren.

Een manier om hier toch enigszins rekening mee te houden is oog te hebben voor voor de hand liggende ‘foutieve’ interpretaties van de data, en deze expliciet te adresseren in metadata, documentatie en andere toelichting. Een voorbeeld zijn de uitgaven voor langdurige zorg (Wlz) naar regio. Worden deze zonder commentaar gepubliceerd dan zullen enkele regio’s er spectaculair uitspringen, en de aandacht trekken in media en publiciteit. Foutieve interpretaties kunnen ten dele voorkómen worden door in de toelichting in te gaan op mogelijke oorzaken van deze verschillen. In het voorbeeld van de Wlz-uitgaven is het voor de hand liggend dat bepaalde regio’s er uit springen omdat ze verpleeghuizen in de regio hebben, waar die in andere gemeenten ontbreken.

6. Voorkom dat anonimiteit van eerder gepubliceerde datasets wordt aangetast

Een onbedoeld effect van publicatie van een geanonimiseerde dataset is dat de anonimiteit van door anderen gepubliceerde datasets over ditzelfde gegeven wordt aangetast en vice versa. Een eenvoudig voorbeeld is de publicatie van eenzelfde ruimtelijk uitsplitsbaar gegeven, maar met een verschillende gebiedsindeling: gebruikers die beide sets combineren kunnen zo gegevens achterhalen voor kleinere ruimtelijke eenheden dan gewenst, door ‘grensverschillen’ tussen beide regio’s uit te buiten. In theorie zouden verschillende gebiedsindelingen moeten worden voorkomen, maar in de praktijk blijkt dat meestal niet haalbaar. Anonimiseren in geval van verschillende gebiedsindelingen kan door de kruising van alle gehanteerde gebiedsindelingen te beschouwen. Dit leidt echter vaak tot veel informatieverlies.

Ook het standaardiseren van productiewijzen binnen en tussen organisaties helpt om tot goede bescherming van gegevens te komen. Vooral het hanteren van standaardclassificaties voor variabelen als regio, beroep en opleidingsniveau kan ervoor zorgen dat anonimiteit van eerder gepubliceerde datasets beter wordt geborgd. Een bruikbare standaardclassificatie is een classificatie waarvan een publieke beschrijving bestaat én waarvan het onderhoud (indien relevant) is geborgd. Voorbeelden zijn de ATC-classificatie voor farmacie en de ICD-classificatie voor ziekte-diagnose (beide door de WHO ontwikkeld en geborgd). Waar geen beschrijving/borging bestaat (denk bijvoorbeeld aan leeftijdsklassen), kan worden verwezen naar normen of standaarden (zoals de diverse leeftijdsklassen die het CBS hanteert).

Er zijn geen algemene richtlijnen te geven. Van ieder gepubliceerd geanonimiseerd gegeven kan de anonimiteit worden aangetast door op moment van publicatie nog niet beschikbare of zelfs maar verzamelde data. Wel is het aan te bevelen binnen de eigen organisatie een goed 'omgevingsbewustzijn' voor dit fenomeen te propageren en voldoende kennis in huis te hebben van de data die andere partijen eerder hebben gepubliceerd of nog gaan publiceren op hetzelfde terrein, en hiermee rekening te houden bij de keuze voor een veiligheidsniveau bij het anonimiseren, en indien nodig data te publiceren op een ander niveau.

7. Houd rekening met toekomstige technische mogelijkheden die anonimiteit kunnen aantasten

Volgens de Algemene Verordening Gegevensbescherming (AVG) zijn gegevens pas echt anoniem (en vallen daarmee niet onder de AVG) als het 'redelijkerwijs' onmogelijk is personen te identificeren, rekening houdend met beperkingen in kosten en tijd, en met de huidige technologie en toekomstige technologische ontwikkelingen[21]. Deze eis heeft niet alleen betrekking op nieuwe open te publiceren datasets. Rekening houden met toekomstige ontwikkelingen betekent ook dat moet worden voorkómen dat publicatie van een dataset de anonimiteit van reeds gepubliceerde datasets (van de eigen organisatie of van anderen) in de toekomst zal aantasten en vice versa. Bijvoorbeeld door nieuwe mogelijkheden voor het combineren van datasets (zie ook Aanbeveling 6). Deze eis stelt organisaties die open data willen publiceren voor een lastige uitdaging. Een eenduidig antwoord is niet mogelijk. Van geval tot geval moet bekeken worden hoe hier het beste mee om kan worden gegaan. Mogelijke manieren hiervoor zijn:

- Structureel overleg tussen partijen die gelijksoortige gegevens publiceren;
- Instellen meldpunt/klankbordgroep waar lastige gevallen besproken kunnen worden;
- Een privacy impact assessment uitvoeren vóórdat open data worden gepubliceerd;
- Iedere 5 jaar bekijken of eerder gepubliceerde open data sets in de huidige context nog steeds niet herleidbaar zijn.

De AVG is nog zo vers dat de praktijk moet gaan uitwijzen wat de consequenties van de eis in de AVG zijn om rekening te houden met toekomstige ontwikkelingen. In ieder geval zouden organisaties moeten kunnen aantonen dat ze hierover nagedacht hebben bij het publiceren van open data.

8. Overweeg mogelijke alternatieven voor geanonimiseerde open data

Het is mogelijk dat het door gebruikers gewenste detailniveau te veel privacyrisico's met zich mee brengt, hetzij volgens de wet, hetzij volgens richtlijnen die bronhouders zelf hanteren. Het is dan voor de beoogde gebruiker niet (altijd) zinvol als de geanonimiseerde data toch, maar met minder detailniveau, open worden gepubliceerd, omdat deze data dan niet voldoende van waarde is voor het beoogde hergebruik. Er kan dan over een alternatieve publicatievorm worden nagedacht, waarbij bijvoorbeeld het deel dat open gepubliceerd kan worden als zodanig wordt gepubliceerd, en voor de meer gevoelige

gegevens een alternatieve publicatievorm wordt gekozen. Mede doordat het begrip *anoniem* in de loop van de tijd een andere betekenis heeft gekregen en doordat het steeds moeilijker is geworden om te 'bewijzen' dat een gegeven anoniem is, winnen andere methoden voor het geschikt maken van data voor hergebruik die voldoen aan de FAIR-principes, aan populariteit als alternatief voor anonimiseren van gegevens.

Hieronder worden enkele alternatieven gegeven aan de hand van voorbeelden met de data uit de gezondheidsenquête die het CBS jaarlijks onder 10.000 Nederlanders houdt. Deze data worden op drie manieren gepubliceerd:

(A) Open data in de vorm van statistische (geaggregeerde) uitkomsten (bijvoorbeeld percentages, totalen, gemiddelden en standaardafwijking) op het open data portaal van CBS (StatLine)[22]. Zie Aanbeveling 4B. Hierbij is het van belang voorwaarden te stellen aan het hergebruik van de data. Voor de data op StatLine geldt dat verveelvoudiging is toegestaan, mits het CBS als bron wordt vermeld. Een andere mogelijke oplossing is een Creative Commons licentie waarin staat onder welke voorwaarden het is toegestaan iemands werk te gebruiken of verspreiden[23].

(B) De individuele records worden als 'partieel geanonimiseerde' data beschikbaar gesteld op basis van contracten met geselecteerde gebruikers als universiteiten en beleidsonderzoekers. Dit gebeurt vaak, maar niet uitsluitend, via het dataportaal DANS van de Koninklijke Academie van Wetenschappen[24]. Partieel anonimiseren houdt in dat directe identificatoren (zoals naam, BSN, telefoonnummer) zijn verwijderd en dat het onthullingsrisico is verkleind door bijvoorbeeld de leeftijd in klassen op te nemen i.p.v. in afzonderlijke jaren (globaal hercoderen) en dat zeldzame combinaties van scores op identificerende variabelen worden onderdrukt. Het onthullingsrisico is dus sterk verkleind maar niet helemaal afwezig. Als iemand de scores van een persoon op vele identificerende variabelen kent, kan hij/zij die persoon soms toch identificeren. Volgens de contractuele voorwaarden is het echter niet toegestaan om op zoek te gaan naar informatie over specifieke individuen. Bovendien mogen de DANS-bestanden volgens de contractuele voorwaarden ook niet worden gekoppeld en er zit ook geen koppelsleutel in de bestanden. De beveiliging van DANS-bestanden is daarmee gedeeltelijk statistisch ('partieel anonimiseren') en gedeeltelijk juridisch ('contract').

(C) De statische beveiliging van de bestanden komt overeen met die van voorbeeld B, maar nu wordt het bestand op dezelfde wijze gepseudonimiseerd als bijvoorbeeld de basisregistratie personen of de doodsoorzakenstatistiek. Dit maakt het mogelijk records te koppelen aan gegevens uit die andere bestanden. Dit gepseudonimiseerde en koppelbare bestand is beschikbaar voor onderzoekers in een beveiligde omgeving bij het CBS of op het eigen instituut van de onderzoeker via een remote access voorziening. Ook hierbij is de toegang gereguleerd via een contract.

Alleen voor bestand (A) zijn de data volledig geanonimiseerd. Bij (B) en met name (C) zijn de data theoretisch te ont-anonimiseren, maar dit wordt afgedekt door de bepalingen in de contracten met gebruikers, waarin uitsluitend publicaties worden toegestaan van anonieme statistische (geaggregeerde) uitkomsten. De meest gevoelige – koppelbare – data onder (C) zijn daarnaast uitsluitend te bewerken in een beveiligde omgeving, waarbij iedere door onderzoekers mee te nemen output wordt gecontroleerd op mogelijke onthulling alvorens die wordt vrijgegeven.

Bij deze voorbeelden afkomstig van het CBS dient aangetekend te worden dat het CBS een geheel eigen systeem heeft dat gebaseerd is op specifieke wetgeving (CBS-wet). De geschetste alternatieven kunnen daardoor niet zonder nadere toets aan privacy wet -en regelgeving worden 'vertaald' naar en toegepast door zorgpartijen.

9. Publiceer minimaal metadata

Het via contracten of in een beveiligde omgeving beschikbaar stellen van microdata of publicatie van tabellen met geaggregeerde data kunnen dus alternatieven zijn voor publicatie van open data. Volgens de FAIR-principes moeten data in ieder geval vindbaar zijn voor potentiële gebruikers, wat betekent dat een metadata-beschrijving gepubliceerd zou moeten worden, ook als niet tot open data publicatie wordt over gegaan, maar data alleen onder voorwaarden voor geselecteerde gebruikers toegankelijk zijn. Metadata zijn data over de data. Het gaat om gegevens die de karakteristieken van bepaalde gegevens beschrijven. Het publiceren van uitsluitend de metadata is te zien als een extreme vorm van anonimiseren, waarbij geen enkel gegeven open wordt gepubliceerd, en data uitsluitend 'vindbaar' zijn en niet 'beschikbaar'. Diverse partijen faciliteren de publicatie van metadata en het is aan te bevelen de metadata-beschrijving van een geanonimiseerde dataset hier onder te brengen. Uiteraard kan een organisatie ook besluiten zelf metadata te publiceren op een eigen website, maar door aan te sluiten bij bestaande metadata-catalogi wordt de vindbaarheid van bestanden voor externe partijen vergroot, en wordt ook de uniformiteit van de metadata-beschrijvingen verhoogd. De keuze voor waar de beschrijving onder te brengen hangt af van het type aangeboden data.

De belangrijkste publicaties van metadata-bestanden die ook open beschrijvingen van 'gesloten' data bevatten zijn:

Open data portaal overheid: https://data.overheid.nl/. Op deze site zijn metadata-beschrijvingen te vinden van, voor publieke taken verzamelde, open data. Er wordt ook doorverwezen naar de datasets zelf.

Zorggegevens.nl: https://www.volksgezondheidenzorg.info/zorggegevens. In de metadatabase Zorggegevens is een overzicht opgenomen wie welke gegevens verzamelt over volksgezondheid en zorg in Nederland, met welk doel, wie dit financiert en waar de gegevens voor gebruikt worden. Het gaat daarbij om in Nederland beschikbare bronnen zoals zorgregistraties, patiëntenregistraties, enquêtes, monitors, langlopende (cohort) onderzoeken en andere onderzoeken waarvan de data beschikbaar zijn voor hergebruik. Deze metadatabase is een initiatief van het ministerie van Volksgezondheid, Welzijn en Sport (VWS) en wordt onderhouden door het RIVM.

CBS-micro-data: https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/microdata-zelf-onderzoek-doen/catalogus-microdata. Het CBS heeft een grote verzameling microdata opgebouwd. Dit zijn data veelal op persoonsniveau die onderling koppelbaar zijn met behulp van een gepseudonimiseerd persoonsnummer. Koppelen kan alleen onder strikte voorwaarden binnen een beveiligde omgeving bij het CBS plaatsvinden, op basis van een contract. Naast het CBS kunnen ook de oorspronkelijke registratiehouders van de data eisen stellen aan hergebruik. Een deel van deze bestanden is afkomstig

uit het domein gezondheid en welzijn. Een klein deel van deze bestanden is ook beschikbaar als onderzoeksbestand bij DANS van de Koninklijke Nederlandse Academie van Wetenschappen (KNAW).

DANS-EASY onderzoeksbestanden: https://dans.knaw.nl/nl. De KNAW biedt een faciliteit voor het deponeren van onderzoeksbestanden. Doel is het delen en hergebruiken van in academisch onderzoek verzamelde data. Van alle gedeponeerde bestanden is een metadata-beschrijving beschikbaar, en onder voorwaarden kunnen datasets worden opgevraagd en hergebruikt.

BBMRI-Biobanken: https://catalogue.bbmri.nl/. BBMRI staat voor 'Biobanking and BioMolecular resources Research Infrastructure The Netherlands', en verenigt alle organisaties die in Nederland biologisch en diagnostisch materiaal verzamelen. Denk daarbij aan bloed- en weefselmonsters, maar ook aan bijvoorbeeld scannerdata. Veel data zijn verzameld als onderdeel van langlopende bevolkingsonderzoeken.

3. Praktijkvoorbeelden

Veel partijen zijn uit eigen beweging al overgegaan tot het open beschikbaar stellen voor hergebruik door derden van geanonimiseerde data uit registers met informatie die zorgverleners routinematig vastleggen, als onderdeel van de zorgverlening of voor declaratie. Hierin staat bijvoorbeeld informatie over welke behandelingen een patiënt ondergaat, welke diagnoses zijn gesteld, welke zorg is verstrekt en de administratieve afhandeling van geleverde zorg. Aansprekende voorbeelden hiervan zijn te vinden bij bijvoorbeeld Vektis, het Zorginsituut, het RIVM, het Nivel, de Nederlandse Zorgautoriteit en het CBS. Op het open data portaal van de overheid zijn inmiddels ruim 700 zorggerelateerde sets gepubliceerd. Zie: https://data.overheid.nl/data/dataset?q=zorg

Het merendeels is afkomstig van het CBS. In veel gevallen gaat het hier echter niet om data die op individueel niveau vrij beschikbaar wordt gesteld. Deze data worden namelijk niet zonder restricties open gepubliceerd. Dat geldt wel voor geaggregeerde data, waarbij het in principe onmogelijk is gemaakt om individuele burgers te identificeren. Hierna volgen vier voorbeelden van hoe Vektis, RIVM, NZa en Nivel omgaan met het al of niet open beschikbaar stellen van data voor hergebruik door derden.

Voorbeeld open data Vektis

Auteurs: Paul Merkx en Dick Johan van der Harst

Inleiding

Vektis verzamelt en analyseert gegevens over de kosten en de kwaliteit van de gezondheidszorg in Nederland en stelt partijen in staat de kwaliteit, toegankelijkheid en betaalbaarheid van de zorg in Nederland te verbeteren. In opdracht van de zorgverzekeraars publiceert Vektis sinds 2014 ook open data.

De Vektis Open Databestanden Zorgverzekeringswet zijn openbaar toegankelijke databestanden met de - onder de basisverzekering - in recente jaren gedeclareerde zorgkosten. Daarbij zijn de zorgkosten onderverdeeld naar de verschillende zorgsoorten die de Zorgverzekeringswet onderscheidt. Daarnaast bevatten de databestanden enkele persoonskenmerken van zorgverzekerden: geslacht, leeftijd en de woonlocatie. Bij de woonlocatie betreft het ofwel de gemeentenaam ofwel de eerste drie cijfers van de postcode. De databestanden hebben een landelijke dekking en bevatten dus gegevens van alle zorgverzekerden. De gegevens zijn echter niet herleidbaar tot individuele personen, zorgverzekeraars of zorgaanbieders.

Gegevensbewerking en waarborging privacy

Bestanden met declaratiegegevens, die Vektis periodiek van de zorgverzekeraars ontvangt, vormen de basis van de open databestanden. Vektis aggregeert de declaratiegegevens tot bestanden met minimaal tien achterliggende personen per datarecord. Indien in een combinatie van persoonskenmerken minder dan tien personen voorkomen, zijn de bijbehorende verzekerdenaantallen en zorgkosten ondergebracht

in een restcategorie zonder gegevens over leeftijd, geslacht en woonlocatie. Het datawarehouse van Vektis bevat geen BurgerServiceNummers (BSN's) of volledige adresgegevens en in de databron voor het maken van de open databestanden is geen gebruikgemaakt van volledige geboortedatums.

Beschikbaarstelling van gegevens

Gegevens worden in CSV-bestandsformaat beschikbaar gesteld op: https://www.vektis.nl/streams/open-data. Voor elk jaar is een apart bestand beschikbaar. Er zijn bestanden beschikbaar per gemeente en op postcode3-niveau. Een bijsluiter geeft uitgebreide achtergrondinformatie over inhoud en totstandkoming van de bestanden[25].

Conclusie en discussie

Bij het bepalen van de inhoud van de open databestanden opereert Vektis steeds in het spanningsveld tussen enerzijds de maatschappelijke roep om transparantie en meer informatie en anderzijds de verantwoordelijkheid om namens de zorgverzekeraars te voldoen aan wetgeving op het gebied van privacy en mededinging. Het is zaak in dit spanningsveld de juiste middenweg te vinden.

In samenwerking met Zorgverzekeraars Nederland is in 2015 een adviesraad open data gevormd waarin experts zitting hebben van onder andere het CBS, het ministerie van VWS en andere partijen uit het zorgveld. Deze adviesraad heeft tot doel om Vektis en de zorgverzekeraars te adviseren over de opzet en inhoud van de open databestanden. De uiteindelijke besluitvorming over de publicatie en inhoud van de open databestanden ligt echter bij de zorgverzekeraars. Het is dan ook van belang dat Vektis in staat is aan de verzekeraars te laten zien dat de te publiceren open data op zodanige wijze beveiligd zijn, dat onthulling van gegevens over personen (of van gegevens over individuele zorgverzekeraars of zorgaanbieders) niet mogelijk is.

Vektis krijgt veelvuldig verzoeken van derden om andere informatie of meer detailinformatie in de open databestanden op te nemen. Zo is er veel vraag naar een meer fijnmazige regio-indeling dan op dit moment beschikbaar is. In grote gemeenten zijn wijken bijvoorbeeld niet te onderscheiden op basis van driecijferige postcodes. Ook zijn er verzoeken om data die betrekking hebben op zorggebruik in plaats van enkel zorgkosten. Om dergelijke bestanden in de toekomst te kunnen publiceren, moeten deze op de juiste wijze statistisch beveiligd worden. Daarbij is het van belang om aan de zorgverzekeraars aan te tonen dat hiermee het risico op onthulling minimaal is en dat aangesloten wordt op algemeen geldende richtlijnen die breed gedragen worden door organisaties die, net als Vektis, open data op het gebied van de gezondheidszorg publiceren.

Voorbeeld open data RIVM

Auteur: Laurens Zwakhals

Inleiding

De Dienst Vaccinvoorziening en Preventieprogramma's (DVP) coördineert de uitvoering van onder meer het Rijksvaccinatieprogramma (RVP). Voor dit programma koopt DVP de benodigde vaccins in. DVP beheert de voorraden en distribueert deze naar de uitvoerders van het programma.

DVP roept de mensen op voor het RVP, registreert alle gegeven vaccinaties op individueel niveau in het 'Praeventis-systeem', adviseert in individuele gevallen, verwijst zo nodig door en monitort de uitvoering. Ook verzorgt DVP scholing voor de uitvoerenden. Door advisering over de uitvoering aan het ministerie van Volksgezondheid, Welzijn en Sport (VWS) draagt DVP bij aan de beleidsvorming over preventieve programma's.

DVP heeft voor deze programma's een verbindende rol tussen individuele burgers, gezondheidsorganisaties en de overheid. Opdrachtgever voor deze programma's is het ministerie van Volksgezondheid, Welzijn en Sport. De regie op het programma wordt gevoerd door het Centrum Infectieziektebestrijding (CIb) van het RIVM.

Over de volgende vaccinaties wordt gerapporteerd:

Bij zuigelingen (0-14 maanden):

- DKTP
- Hib
- BMR
- Meningokokken C
- Hepatitis B
- Volledige deelname aan alle bovengenoemde vaccinaties op de leeftijd van 2 jaar

Bij kleuters (4 jaar):

- D(K)TP

Schoolkinderen (9 jaar)

- DTP
- BMR

Adolescente meisjes (12 jaar)

- HPV

Voor al deze groepen wordt jaarlijks de populatie, het aantal gevaccineerde kinderen en de vaccinatiepercentages gerapporteerd voor heel Nederland, per GGD-regio en per gemeente

Gegevensbewerking en waarborging privacy

Het RIVM rapporteert jaarlijks over de vaccinatiegraad van het RVP. Daarvoor worden speciale verzoeken gedaan aan *Praeventis*. Er worden vooraf gedefinieerde queries uitgevoerd op de database.

Vervolgens wordt bekeken of niet over te kleine populaties wordt gerapporteerd. Voor deze dataset kan dat voorkomen in het geval van kleine gemeenten. In die gevallen kan sprake zijn van onthullingsrisico. Om dat te bepalen maakt het RIVM gebruik van de procedure die het CBS hiervoor hanteert. Het CBS past statistische beveiliging toe om dergelijke onthulling te voorkomen. De parameters die het CBS daarbij hanteert zijn niet openbaar, maar de gebruikte methoden wel (zie bijvoorbeeld hoofdstuk 4 over de beveiliging van frequentietabellen in het CBS rapport Statistische Beveiliging[26].

Beschikbaarstelling van gegevens

De uitkomsten van de bevraging van *Praeventis* worden verwerkt in de jaarlijkse rapportages. Een voorbeeld hiervan is hier te vinden:
https://www.rivm.nl/Documenten_en_publicaties/Wetenschappelijk/Rapporten/2018/Juni/Vaccinatiegraad_en_jaarverslag_Rijksvaccinatieprogramma_Nederland_2017

Sinds enige jaren worden de gegevens ook als open data in Excel beschikbaar gesteld. En sinds 2017 ook als open data computer-leesbaar op:
https://statline.rivm.nl/portal?_catalog=RIVM&_la=nl&tableId=50033NED&_theme=61

Bij de publicatie van open data wordt een *Creative Commons* licentie gebruikt: CC-BY SA. Dat betekent dat het anderen toegestaan is om de gegevens te kopiëren, distribueren, vertonen, en op te voeren, en om afgeleid materiaal te maken dat op deze gegevens gebaseerd is – maar uitsluitend met naamsvermelding van de bronhouder. Daarbij wordt als voorwaarde gesteld dat het afgeleide materiaal onder dezelfde licentie wordt vrijgeven als het originele werk (zie ook http://creativecommons.nl/uitleg/).

Conclusie en discussie

Het RIVM is voor de privacy procedure nu nog afhankelijk van het CBS. Het RIVM wil hierover in de toekomst graag zelfstandig opereren waarbij het de beveiligingscriteria van het CBS volgt.

Voorbeeld open data NZa

Auteurs: Dick Odijk

Inleiding

Dit voorbeeld laat zien dat voortschrijdend inzicht, veranderende wetgeving en veranderende omstandigheden in de rol van de organisatie, kunnen leiden tot het besluit om een jaarlijkse rapportage te beëindigen.

De LZV (Landelijk Zorgsysteem voor Veteranen) is een ketenorganisatie binnen het ministerie van defensie die zorgt voor begeleiding en 1e lijns opvang van veteranen met (mogelijk) psychische problemen die samenhangen met eerdere uitzendingen. De LZV heeft behoefte aan meer gegevens over zorggebruik van de veteranen dan dat de organisatie zelf tot zijn beschikking heeft in de eigen database 'De Basis'. Daarvoor leverde de LZV hun gegevensbestand via ZorgTTP gepseudonimiseerd aan de NZa die de gegevens verrijkte met meer informatie over zorggebruik (DBC-Informatie).

Na koppeling met GGZ DBC-Informatie werd er over de gehele set van data van de veteranen (1e lijns informatie + DBC-informatie) een rapportage gemaakt. Hierin werd onder meer opgenomen:

- Gemiddelde wachttijd in dagen van melding tot 1e gesprek;
- Gemiddelde wachttijd in dagen van melding tot 1e gesprek waarbij sprake is van een vervolg DBC;
- Afsluiting DBC na alleen pré-intake/intake/diagnostiek/crisisopvang;
- DBC in onderling overleg beëindigd zorgtraject/ patiënt uitbehandeld;
- Reden voor afsluiting DBC bij patiënt / niet bij behandelaar (z.g. DROP-OUT);
- Verblijfsduur in de klinische keten.

Daarbij werden uitsplitsingen gemaakt naar:

- Woonregio
- Uitzendgebied
- Geslacht
- Zorgsoort (1e, 2e lijn)
- Zorginstelling (indien daarvoor goedkeuring is verleend)

Gegevensbewerking en waarborging privacy

Jaarlijks leverde de LZV via ZorgTTP een bestand met daarin de individuele gegevens van de veteranen, geregistreerd in De Basis.

- Algemene persoonsgegevens: geboortejaar, geslacht;
- Specifieke persoonsgegevens: BSN-nummer, door ZorgTTP omgezet naar (DIS)pseudoniem;
- Contactgegevens 1e lijn;
- Uitzending(en).

Met behulp van het BSN-pseudoniem is een individuele koppeling mogelijk tussen de 1e lijns gegevens van de veteraan en de (mogelijk aanwezige) DBC-gegevens uit DIS. Na koppeling werd er over de gehele set van data van de veteranen (1e lijns informatie + DBC-informatie) een rapportage gemaakt.

Problematiek:

1. Koppeling van 1e lijns individuele registraties aan individuele 2e lijns DBC-gegevens. Instrument: BSN-pseudoniem;
2. Hoewel er meestal sprake was van minimaal 3 veteranen waarop de gemiddelden uit de rapportage gebaseerd waren, waren 1 of 2 waarnemingen in de groep ook mogelijk. Zeker bij rapportages per 2e lijns zorginstelling kwam dit geregeld voor;
3. Rapportage van cijfers op niveau van een individuele zorginstelling.

Beschikbaarstelling van gegevens

De uitlevering van deze rapportage is na 2016 stopgezet.

Conclusie en discussie

De laatste keer dat deze rapportage voor de LZV is gepubliceerd is in het voorjaar van 2016:

- Op grond van de individuele opdracht aan destijds DBCOnderhoud/DIS;
- Data 1e lijn afkomstig uit de eigen registratie;
- Toestemming van alle betrokken 2e lijns zorginstellingen;
- Meetwaarden buiten de 2e lijns zorginstellingen zijn bijna altijd gebaseerd op 3 of meer waarnemingen;
- Bij minder dan 3 waarnemingen werden geen meetwaarden opgenomen. Individuele personen waren hierdoor onmogelijk traceerbaar.

In 2015 is DBC-Onderhoud opgenomen binnen de NZa; waarbij deze de publieke taken en verantwoordelijkheden van DBC-Onderhoud heeft overgenomen. In eerste instantie inclusief de jaarlijkse rapportage aan de LZV.

Eind 2016 is echter besloten op grond van de aangescherpte eisen van de Autoriteit Persoonsgegevens, in samenhang met de rol van de NZa (die anders is als die van DBCOnderhoud/DIS in het verleden) om af te zien van de verdere doorontwikkeling en uitlevering van deze rapportage.

Voorbeeld open data Nivel

Auteur: Robert Verheij

Inleiding

Nivel Zorgregistraties is één van de onderdelen van de onderzoeksinfrastructuur waar het Nivel verantwoordelijk voor is (naast consumentenpanels en beroepskrachtenregistraties). Binnen Nivel Zorgregistraties worden gegevens verzameld over zorg en gezondheid die routinematig worden bijgehouden in elektronische patiëntendossiers van zorgverleners. Het gaat hierbij primair om gecodeerde velden, dus niet om vrije tekst. Het gaat om aan de bron, onomkeerbaar gepseudonimiseerde gegevens over gezondheidsproblemen, geneesmiddelengebruik, verwijzingen, behandelingen. Geboortedatum wordt voor verzending naar het Nivel omgezet in geboortekwartaal. De vier cijfers van de postcode komen mee, evenals geslacht. Het gaat om gegevens van huisartsen (met in totaal 1,7 miljoen ingeschreven patiënten), huisartsenposten (met een verzorgingsgebied van in totaal ruim 11 miljoen inwoners) en daarnaast gegevens van kleinere aantallen fysio- en oefentherapeuten, diëtisten, logopedisten.

Met zorgpraktijken is overeengekomen dat individuele praktijken niet zonder toestemming herkenbaar zullen zijn in rapportages en ook niet in zogenaamde 'open data'.

Doel van de gegevensverzameling is het volgen van ontwikkelingen in gezondheid en zorggebruik van de Nederlandse bevolking en daarover rapporteren op www.nivel.nl/zorgregistraties. Daarnaast worden de gegevens veelvuldig gebruikt voor verder wetenschappelijk onderzoek. Zo krijgt Nivel Zorgregistraties jaarlijks meer dan 50 vragen om gegevens voor verder onderzoek.

Gegevensbewerking en waarborging privacy

De gegevens worden aan de bron onomkeerbaar gepseudonimiseerd, dus voor verzending naar het Nivel. Basis voor de pseudonimisering is het Burger Service Nummer (BSN). Het BSN zelf verlaat de zorgpraktijk niet.

Er is sprake van éénwegpseudonimisering. Daarbij wordt gebruik gemaakt van de diensten van het bedrijf ZorgTTP. De omgeving waarin de gegevens bij het Nivel worden opgeslagen is NEN7510 en ISO 27001 gecertificeerd. Dit zijn de huidige normen voor gegevensbescherming. De procedure met betrekking tot pseudonimiseren staat beschreven in een artikel in het boek Data protection on the move [27].

De gegevens worden primair gebruikt voor het genereren van kerncijfers op de website van Nivel Zorgregistraties (http://www.nivel.nl/zorgregistraties) maar ze worden ook veelvuldig gebruikt voor verder wetenschappelijk onderzoek.

Er is een governance document waarin staat beschreven hoe onderzoeksaanvragen en gegevensaanvragen in behandeling worden genomen. Basisprincipe is dat deelnemende zorgpraktijken de zeggenschap over het gebruik van hun gegevens delegeren aan 'hun' koepelorganisatie. Onderzoek waarbij ook de (vier cijfers van de) postcode nodig zijn, of waarbij koppelingen moeten worden gemaakt, worden voorgelegd aan een privacy-commissie. Voor iedere aanvraag wordt een apart projectnummer

aangemaakt en worden ‘op maat’ bestanden klaargezet. Zo voorkomen we misinterpretaties, houden we zicht op wat er met de gegevens gebeurt, en zijn de geleverde gegevens naspeurbaar om het onderzoek eventueel mee te repliceren. Kwaliteitscontroles en selecties worden per gegevensaanvraag uitgevoerd, omdat ieder onderzoek weer andere eisen stelt aan de benodigde gegevens.

De gegevens die worden uitgeleverd worden voor uitgifte anoniem gemaakt. Om weer terug te kunnen naar de Nivel database, worden records tevens voorzien van een pseudoniem. Dat gebeurt voor iedere gegevensaanvraag opnieuw, zodat onderlinge koppeling van verschillende gegevensaanvragen, met de daaraan verbonden privacyrisico’s, onmogelijke wordt gemaakt. Hier gaat het om tweewegpseudonimisatie (in tegenstelling tot de eerste pseudonimisatieslag via ZorgTTP), zodat de uitgeleverde gegevens ook weer kunnen worden teruggevonden in de Nivel Zorgregistraties database en op die manier de integriteit van de data kan worden aangetoond. Het pseudonimisatieproces staat elders meer uitgeschreven[28].

Beschikbaarstelling van gegevens

Rapportages in de vorm van geaggregeerde kengetallen zijn beschikbaar als open data, bijvoorbeeld incidenties en prevalenties van ziekten en aandoeningen in de huisartsenpraktijk. Zie bijvoorbeeld https://www.nivel.nl/nl/NZR/incidenties-en-prevalenties

Gegevens van Nivel Zorgregistraties kunnen worden aangevraagd voor verder onderzoek. Dit gebeurt veelvuldig. Er gelden daarvoor wel voorwaarden, die zijn vastgelegd in de governance structuur. Tevens worden de principes van FAIR data nagestreefd. Gegevens die voor een bepaald onderzoek zijn uitgeleverd, zijn opvraagbaar voor andere onderzoekers, zodat studies kunnen worden gerepliceerd. Sommige tijdschriften eisen dat gegevensbestanden opvraagbaar zijn bij een onafhankelijke instelling. In die gevallen slaat het Nivel de metadatabestanden op bij DANS.

Gegevens uit Nivel Zorgregistraties kunnen met behulp van pseudoniemen gekoppeld worden aan andere databronnen. Vaak fungeert het CBS daarbij als spil. De beschikbaarheid van de gemeentelijke basisadministratie bij het CBS is daarbij een belangrijke factor. Gegevens dienen dan wel binnen de beveiligde (virtuele) omgeving van het CBS geanalyseerd te worden.

Conclusie

Geaggregeerde gegevens uit Nivel Zorgregistraties worden in de vorm van kengetallen openbaar gepubliceerd. Gegevensbestanden die worden aangevraagd voor verder onderzoek worden zo veel mogelijk anoniem gemaakt. Het blijven doorgaans echter persoonsgegevens, en deze worden niet openbaar gepubliceerd. Bij uitgifte worden de gegevens een tweede keer gepseudonimiseerd om te voorkomen dat verschillende gegevensbestanden met elkaar gecombineerd kunnen worden. De gegevens van Nivel Zorgregistraties zijn als FAIR te kenschetsen en daarmee geschikt voor hergebruik.

Begrippenlijst

Anonimiseren

Anonimiseren of anoniem maken wil zeggen het onherroepelijk verwijderen van gegevens uit databestanden die de identiteit van een persoon kunnen onthullen. In Europese wetgeving (algemene verordening gegevensbescherming) is hier aan toegevoegd dat het terugdraaien van anonimisering in het heden, maar ook in de toekomst met meer geavanceerde technieken 'redelijkerwijs' onmogelijk moet zijn. "Om uit te maken of van middelen redelijkerwijs valt te verwachten dat zij zullen worden gebruikt om de natuurlijke persoon te identificeren, moet rekening worden gehouden met alle objectieve factoren, zoals de kosten van en de tijd benodigd voor identificatie, met inachtneming van de beschikbare technologie op het tijdstip van verwerking en de technologische ontwikkelingen." Bij anonimiseren wordt geen koppelsleutel meegeleverd zodat de geanonimiseerde data nooit meer gekoppeld kan worden aan datasets die meer data over de betreffende persoon of het huishouden bevatten.

Open data

Data worden beschouwd als open data als ze openbaar beschikbaar zijn, en er geen auteursrecht of andere rechten van derden op berusten. Bij voorkeur wordt binnen de open data uitsluitend gewerkt met zogeheten 'open standaarden' om gegevens vast te leggen, die eveneens openbaar gepubliceerd zijn en zonder beperkingen gebruikt mogen worden. Ook zijn open data bij voorkeur computer-leesbaar en doorzoekbaar voor zoekmachines. Voor een uitgebreidere definitie zie het Dataportaal van de Nederlandse overheid op https://data.overheid.nl/

Binnen het overheidsdomein is in de 'Wet hergebruik overheidsinformatie' (http://wetten.overheid.nl/BWBR0036795/2016-10-01) afgesproken dat data die verzameld is in het kader van een publieke taak en bekostigd zijn met publieke middelen als open data dienen te worden gepubliceerd door de partijen die de data vastleggen.

Pseudonimiseren

Pseudonimiseren is het verwerken van persoonsgegevens op zodanige wijze dat de persoonsgegevens niet meer aan een specifieke betrokkene kunnen worden gekoppeld zonder dat er aanvullende gegevens worden gebruikt, mits deze aanvullende gegevens apart worden bewaard en technische en organisatorische maatregelen worden genomen om ervoor te zorgen dat de persoonsgegevens niet aan een geïdentificeerde of identificeerbare natuurlijke persoon worden gekoppeld. Deze definitie is ontleend aan de Europese wetgeving (algemene verordening gegevensbescherming).

Bij pseudonimiseren wordt in technische zin direct identificerende persoonsinformatie (bijvoorbeeld naam adres, burgerservicenummer) vervangen door een code. Het voordeel van pseudonimiseren is dat bestanden die op dezelfde wijze zijn gepseudonimiseerd zijn, nog steeds te koppelen zijn. Pseudonimiseren is niet voldoende om te anonimiseren, omdat theoretisch identificatie mogelijk blijft voor degene die over de pseudonimiserings-sleutel beschikt.

Meer lezen

Juridische aspecten

Actuele informatie over de Europese privacy wetgeving en de implementatie hiervan in Nederland is te vinden op de website van de Autoriteit Persoonsgegevens.
https://autoriteitpersoonsgegevens.nl/nl/onderwerpen/europese-privacywetgeving

De officiële Nederlandse tekst van de algemene verordening gegevensbescherming (verordening (EU) 2016/679 van het Europese parlement en de raad van 27 april 2016): http://eur-lex.europa.eu/legal-content/NL/TXT/PDF/?uri=CELEX:32016R0679&rid=1

Wet hergebruik van overheidsinformatie (Who): http://wetten.overheid.nl/BWBR0036795/2016-10-01

Technische aspecten

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte Nordholt, K. Spicer and P.P. de Wolf, 2012. Statistical Disclosure Control. In: Wiley Series in Survey Methodology. Wiley, Chichester, United Kingdom, 2012.

Hundepool A, De Wolf P-P. Statistische beveiliging. Centraal Bureau voor de Statistiek, Den Haag/Heerlen, 2010. https://www.cbs.nl/nl-nl/onze-diensten/methoden/statistische-methoden/output/output/statistische-beveiliging

Specifieke software voor beveiliging van statistische tabellen (Tau-ARGUS) en de bijbehorende manual zijn te vinden op: http://neon.vb.cbs.nl/CASC/tau.htm

De nieuwste versies van verschillende softwarepakketten voor Statistische Beveiliging met vragen en antwoorden zijn te vinden op: https://github.com/sdcTools/

Referenties

[1] Wet hergebruik van overheidsinformatie (Who) http://wetten.overheid.nl/BWBR0036795/2016-10-01
[2] Handleiding Wet hergebruik van overheidsinformatie. https://open-overheid.nl/wp-content/uploads/2016/01/WEB_88737_handleiding_A5.pdf
[3] Wet op het Centraal bureau voor de statistiek. https://wetten.overheid.nl/BWBR0015926/2018-07-28
[4] Learning Healthcare Project. Learning Healthcare Systems http://www.learninghealthcareproject.org
[5] Delaney BC, Peterson KA, Speedie S, Taweel A, Arvanitis TN, Hobbs FD. Envisioning a learning health care system: the electronic primary care research network, a case study. Annals of family medicine. 2012;10(1):54-9.
[6] Friedman C, Rubin J, Brown J, Buntin M, Corn M, Etheredge L, et al. Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. Journal of the American Medical Informatics Association : JAMIA. 2015;22(1):43-50.
[7] Kamerbrief Data laten werken voor gezondheid. https://www.rijksoverheid.nl/documenten/kamerstukken/2018/11/15/kamerbrief-over-data-laten-werken-voor-gezondheid
[8] ZonMw. Fair Data. https://www.zonmw.nl/nl/over-zonmw/toegang-tot-data/fair-data/
[9] Van Loenen B, Welle Donker FM, Ploeger HD. RIVM open data.
Kenniscentrum open data Faculteit Bouwkunde, Technische Universiteit Delft. 2016.
https://pure.tudelft.nl/portal/files/4909863/2016_RIVM_open_data.pdf
[10] Peersman C, Daelemans W, Van Vaerenbergh L. Predicting age and gender in online social networks. Proceedings of the 3rd international workshop on Search and mining user-generated contents. ACM, 2011. Te downloaden via: http://www.clips.ua.ac.be/~walter/papers/2011/pdv11.pdf
[11] Algemene verordening gegevensbescherming (verordening (EU) 2016/679 van het Europese parlement en de raad van 27 april 2016) http://eur-lex.europa.eu/legal-content/NL/TXT/PDF/?uri=CELEX:32016R0679&rid=1
[12] Groep gegevensbescherming artikel 29. Advies 5/2014 over anonimiseringstechnieken. 0829/14/NL. WP 216.
[13] Algemene verordening gegevensbescherming (verordening (EU) 2016/679 van het Europese parlement en de raad van 27 april 2016) http://eur-lex.europa.eu/legal-content/NL/TXT/PDF/?uri=CELEX:32016R0679&rid=1
[14] Vrij Nederland 2014. 'Als de overheid bijna niks geheim houdt' https://www.vn.nl/als-de-overheid-bijna-niks-meer-geheim-houdt/
[15] Martijn M, Tokmetzis D. Je hebt wél iets te verbergen. Over het levensbelang van privacy. de Correspondent Bv, 2016
[16] Hundepool A, De Wolf P-P. Statistische beveiliging. Centraal Bureau voor de Statistiek, Den Haag/Heerlen, 2010. https://www.cbs.nl/nl-nl/onze-diensten/methoden/statistische-methoden/output/output/statistische-beveiliging
[17] Groep gegevensbescherming artikel 29. Advies 5/2014 over anonimiseringstechnieken. 0829/14/NL. WP 216.
[18] Hundepool A, De Wolf P-P. Statistische beveiliging. Centraal Bureau voor de Statistiek, Den Haag/Heerlen, 2010. https://www.cbs.nl/nl-nl/onze-diensten/methoden/statistische-methoden/output/output/statistische-beveiliging
[19] τ-ARGUS home page. http://neon.vb.cbs.nl/CASC/tau.htm
[20] τ-ARGUS home page. http://neon.vb.cbs.nl/CASC/tau.htm
[21] Algemene verordening gegevensbescherming (verordening (EU) 2016/679 van het Europese parlement en de raad van 27 april 2016) http://eur-lex.europa.eu/legal-content/NL/TXT/PDF/?uri=CELEX:32016R0679&rid=1
[22] StatLine. https://opendata.cbs.nl/statline/#/CBS/nl/
[23] Uitleg bij de Creative Commons-licenties. http://creativecommons.nl/uitleg/
[24] DANS. https://dans.knaw.nl/nl
[25] Bijsluiter Vektis Open Databestanden Zorgverzekeringswet. Te downloaden via https://www.vektis.nl/streams/open-data
[26] Hundepool A, De Wolf P-P. Statistische beveiliging. Centraal Bureau voor de Statistiek, Den Haag/Heerlen, 2010. https://www.cbs.nl/nl-nl/onze-diensten/methoden/statistische-methoden/output/output/statistische-beveiliging
[27] Kuchinke W, Ohmann C, Verheij RA, Van Veen EB, Delaney B. Development Towards a Learning Health System—Experiences with the Privacy Protection Model of the TRANSFoRm Project. In: Gutwirth S, Leenes R, De Hert P, editors. Data Protection on the Move Current Developments in ICT and Privacy/Data Protection. Issues in Privacy and Data Protection: Springer; 2016.

[28] Kuchinke W, Ohmann C, Verheij RA, van Veen EB, Arvanitis TN, Taweel A, Delaney BC. A standardised graphic method for describing data privacy frameworks in primary care research using a flexible zone model. Int J Med Inform. 2014;83(12):941-57.