ZorgTTP
Privacy & vertrouwen

Factsheet pseudonimisatie SFK cohorten

1. Inleiding

Deze factsheet beschrijft in hoofdlijnen hoe pseudonimisatie ten behoeve van wetenschappelijke onderzoeken (cohort studies) met data uit de SFK databank (hierna: SFK cohorten) plaatsvindt. We beschrijven de uitgangspunten voor en procestappen van de door Stichting ZorgTTP ontwikkelde systematiek voor het pseudonimiseren van persoonsgegevens.

Stichting ZorgTTP is sinds 2007 actief als Trusted Third Party (TTP) en is gespecialiseerd in het ondersteunen van organisaties bij het op passende wijze beschermen van privacygevoelige informatie ten behoeve van beleids- en onderzoeksdoeleinden. ZorgTTP biedt diensten aan op het gebied van privacybeschermende maatregelen waaronder Privacy Enhancing Technologies (PETs) en advieswerkzaamheden ten behoeve van (op te richten) datasamenwerkingen. Stichting ZorgTTP werkt hierbij met inachtneming van de geldende wet- en regelgeving en volgens de stand der techniek. Bij Stichting ZorgTTP vindt geen opslag van (gepseudonimiseerde) gegevens plaats. Tot de Opdrachtgevers van ZorgTTP behoren naast de SFK, onder andere, de Nederlandse zorgautoriteit(NZa), het Ministerie van VWS, Zorginstituut Nederland (ZIN) en alle Universitaire Medische Centra (UMC's).

2. Uitgangspunten

Pseudonimisatie is een maatregel die kan worden ingezet ter bescherming van persoonsgegevens in grootschalige gegevensverwerkingen. Door de (directe) herleidbaarheid van de identificerende persoonsgegevens te beperken wordt de privacy van de betrokkenen beschermd. Pseudonimiseren is niet hetzelfde als anonimiseren, waarbij opgemerkt dient te worden dat het anonimiseren van data niet alleen zeer complex is maar dat voor het bestempelen van data als anoniem ook een hoge barrière dient te worden overwonnen. Pseudonimisatie wordt ingezet in combinatie met andere beschermende maatregelen zoals functiescheiding en beveiligingsmaatregelen.

De kern van pseudonimiseren is het omzetten van direct herleidbare gegevens zoals de naam of unieke identificerende nummers naar één of meerdere codes; de pseudoniemen.

3. Onomkeerbare pseudonimisatie

In deze factsheet wordt het proces van onomkeerbaar pseudonimiseren ten behoeve van SFK cohorten beschreven. Hierbij wordt pseudonimiseren gedefinieerd als *het onomkeerbaar omzetten van een persoonsgegeven naar een niet tot de oorspronkelijke persoon terug te herleiden unieke code*.

4. Essentie

De omzetting verloopt in een aantal stappen waarbij het cruciaal is dat één van deze stappen bij een zogenaamde Trusted Third Party (TTP) wordt uitgevoerd. De bij de TTP uitgevoerde stap is geheim voor zowel de aanbieder als de afnemer van de gegevens in de pseudonimisatieketen. Op deze wijze kan de relatie tussen pseudoniem en persoonsgegeven zowel technisch als organisatorisch worden verbroken. Na pseudonimisatie is het niet langer mogelijk om via het aangemaakte pseudoniem terug te gaan naar de direct identificerende gegevens behorende bij de natuurlijke persoon waarop het pseudoniem betrekking heeft. Het is wel mogelijk om met dezelfde input tot hetzelfde pseudoniem te komen. Op deze wijze kunnen individuen over tijd en plaats gevolgd worden ten behoeve van beleid en onderzoek zonder dat hiervoor direct identificerende gegevens beschikbaar komen buiten de omgeving waarbinnen ze zijn vastgelegd.

5. Procesverloop

Voor het pseudonimiseren van bestanden die persoonsgegevens bevatten ten behoeve van SFK cohorten, heeft ZorgTTP een pseudonimisatieplatform ontwikkeld. Dit omvat naast een aantal software modules ook technische en organisatorische voorzieningen.

Het pseudonimisatieproces voor SFK cohorten bestaat uit de volgende stappen:

1. De onderzoeker biedt een bestand aan dat voldoet aan de vooraf gedefinieerde en overeengekomen berichtspecificaties. Het bestand bevat persoonsgegevens en een SecureID (een betekenisloosnummer aan de hand waarvan een record kan worden onderscheiden, ook wel transactieID genoemd);
2. Het bestand wordt bij en door de databron lokaal verwerkt met de door ZorgTTP beschikbaar gestelde verzendsoftware;
3. De verwerking bestaat uit het:
    a. Omzetten van de persoonsgegevens tot een pre-pseudoniem (d.m.v. een hash).
    b. Beveiligen van de te verzenden gegevens.
4. Hierna volgt transport via een beveiligde internetverbinding naar ZorgTTP;
5. ZorgTTP voert op de centrale verwerkingsomgeving een tweede bewerking uit op de ontvangen pre-pseudoniemen. Voor SFK wordt een specifieke encryptiesleutel gebruikt. Uitkomst van deze bewerking is een definitief pseudoniem dat niet langer algoritmisch kan worden omgezet naar de oorspronkelijke input;
6. Vervolgens worden de gepseudonimiseerde gegevens vrijgegeven om te worden opgehaald door de SFK met hun ontvangstmodule.
7. ZorgTTP heeft tijdens het pseudonimisatieproces geen toegang tot de gegevens. Dit is beveiligd en kan enkel door de ontvangende partij worden ontsleuteld middels de ontvangstmodule.
8. SFK kan vervolgens op basis van de matchende pseudoniemen in de SFK database de bijbehorende gegevens over geneesmiddelengebruik vinden en buiten ZorgTTP om beschikbaar stellen aan de onderzoekers.

Tijdens het pseudonimisatieproces wordt een gelaagd model van beveiliging gehanteerd:

1. Pseudonimisatie op recordniveau;
2. Versleuteling op bestandsniveau;
3. Transportbeveiliging;
4. Controle & autorisatie van afzender van verzonden bestand middels certificaat.

Voordeel van deze gelaagde aanpak is dat als één laag doorbroken wordt, er nog andere beveiligingslagen resteren om de gegevens te beschermen.

Bijlage 1: Aanleverspecificatie

CSV bestand

De PVM SFK leest de te pseudonimiseren gegevens en het SecureID uit een bestand in CSV formaat met de volgende kenmerken

- Het scheidingsteken is een punt komma (;);
- De namen van de variabelen staan op de eerste regel (de z.g.n. headers);
- **De variabele 'Secureid' moet altijd gevuld zijn;**
- De variabelen Uzovi en Verzekerdenummer zijn optioneel;
- Er wordt gebruik gemaakt van de Windows -tekenset (CP1252).

Indien gebruik wordt gemaakt van Microsoft Excel, gebruik dan de optie 'Opslaan als' en kies dan voor 'CSV (gescheiden door lijstscheidingsteken)(*.csv)'. Om het bestand in te zien, kan het beste gebruik worden gemaakt van een tekst-editor zoals Notepad++. Bij het openen van het CSV-bestand in Excel kan het bestand ongewenst automatisch worden aangepast, voorloopnullen kunnen bijvoorbeeld wegvallen.

Voorbeeld

```
Secureid;Birthdate;Name;Initials;Gender;Postalcode;Uzovi;Verzekerdenummer
1;03-09-1957;Hoogvliet;K;m;8745EE;3032;58748925
```

Specificatie variabelen

In onderstaande tabel zijn de specificaties van de variabelen uitgewerkt.

Variabele	Specificatie	Voorbeeld
Secureid	Alfanumerieke waarde	1
Birthdate	dd-mm-eejj	03-09-1957
Name	Geboortenaam	Hoogvliet
Initiaal	Lengte: 1 letter	K
Gender	Toegestaan zijn M, m, V, v	m
Postalcode	NNNNAA, NNNN AA	8745EE
Uzovi (optioneel)	NNNN	3032
Verzekerdenummer (optioneel)	Numerieke waarde, max 15 lang	58748925

Variabele	Toelichting
Secureid	Het SecureID wordt gebruik voor de terugkoppeling van de resultaten van SFK aan de onderzoeker. Het is van belang dat er een betekenisloos nummer wordt gebruikt.
Birthdate	De geboortedatum van de patiënt.
Name	De geboortenaam van de patiënt.
Initiaal	Het initiaal van de patiënt
Gender	Het geslacht van de patiënt.
Postalcode	De postcode van de patiënt
Uzovi	De UZOVI-code van de verzekeraar van de patiënt.
Verzekerdenummer	Het verzekerdenummer van de patiënt.

Geboortenaam
Bij voorkeur wordt de geboortenaam zonder voorvoegsels aangeleverd. De PVM SFK zal de naam normaliseren voorafgaand aan het aanmaken van pseudoniemen. Normalisatie wordt uitgevoerd om de kans op een match te maximaliseren door registratie artefacten te corrigeren. Een paar voorbeelden:

- Conversie van ij naar y;
- Omzetten van diakrieten.

Initialen
Voor de pseudonimisatie van gegevens wordt gebruik gemaakt van de eerste initiaal. Indien er onduidelijkheid is over de initiaal – bijvoorbeeld in het geval van roepnaam Hans en doopnaam 'Johannes' – kan de onderzoeker ook 2 identieke records aanleveren.

Bestandsnaam

De volgende bestandsnaamconventie wordt gehanteerd:
SFK_cohort_[ID afzender]_[eejjmmdd]_[nr].csv

Een voorbeeld van een toegepaste bestandsnaamconventie:
sfk_cohort_UMCZ_20200120_01.csv

Variabelen (namen & controles)

De PVM SFK start met de validatie van de aanwezige kolomlabels, elke afwijking ten opzichte van de specificatie zal leiden tot afkeur van het bestand: foutcode 1200. Het bestand wordt niet verwerkt.

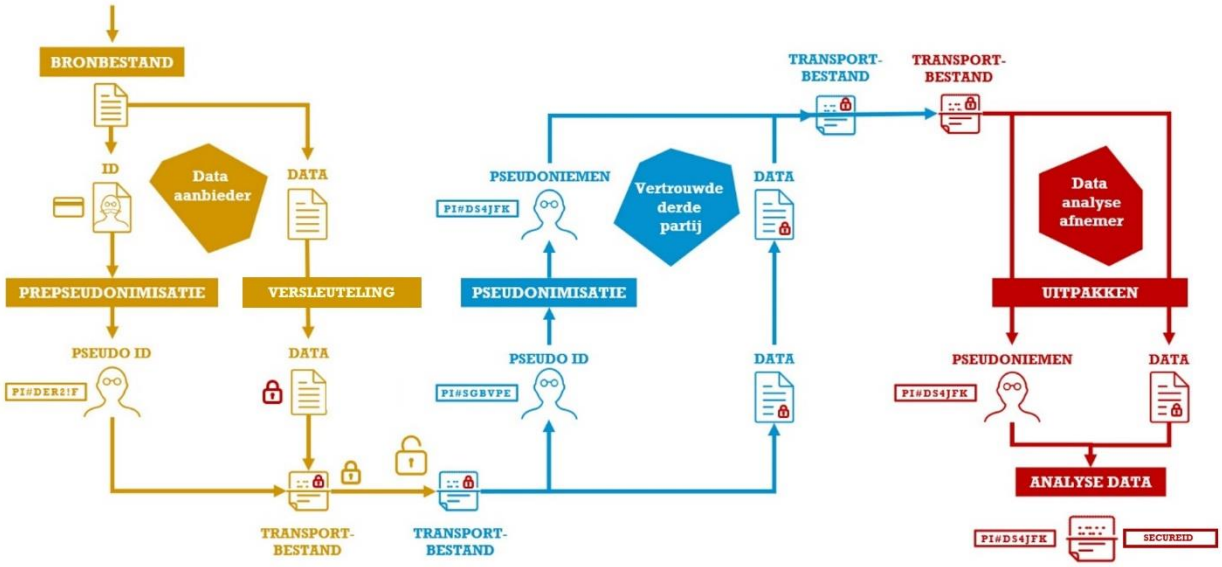
Variabele naam	Foutcode	Omschrijving controle	Effect
	1200	Afwijking in eerste regel CSV	Fatale fout
	1201	Afwijkend aantal kolommen. Elke regel van het bestand moet evenveel kolommen bevatten.	Fatale fout
Name	3011	Naam is leeg	Waarschuwing
	3012	Naam bevat niet toegestane tekens	Waarschuwing
Initials	3021	Voorletter veld is leeg	Waarschuwing
	3022	Voorletters bevatten niet toegestane tekens	Waarschuwing
Birthdate	3031	Geboortedatum is leeg	Waarschuwing
	3033	Geboortedatum ligt in de toekomst	Waarschuwing
	3034	Geboortedatum voldoet niet aan het patroon 31-12-1970 (of voor 1880)	Waarschuwing
Gender	3041	Geslacht is leeg	Waarschuwing
	3042	Geslacht onbekende code	Waarschuwing
Postalcode	3050	Postcode ontbreekt	Waarschuwing
	3051	Postcode is leeg	Waarschuwing
	3052	Postcode voldoet niet aan patroon 1000AA	Waarschuwing
Uzovi	3061	UZOVI is leeg	Waarschuwing
	3062	UZOVI voldoet niet aan patroon 4321	Waarschuwing
Verzekerdenummer	3071	Verzekerdenummer is leeg	Waarschuwing
	3072	Verzekerdenummer voldoet niet aan numeriek tot 15 lang	Waarschuwing

Lege of niet valide velden resulteren in een validatiemelding. Na afronding van de verwerking genereert de PVM een kwaliteitsrapport, waarin alle geconstateerde validaties en bestandsfouten worden getoond. Bepaalde pseudoniemen worden niet aangemaakt als de voor die pseudoniemen benodigde input leeg of niet-valide is. Ontbreken bijvoorbeeld alleen het UZOVI en het verzekerdenummer, dan wordt alleen het V-pseudoniem niet aangemaakt. De overige pseudoniemen zullen daarentegen gewoon worden aangemaakt.

Bijlage 2: Pseudonimisatieketen SFK cohort

Een pseudonimisatieketen is een klantspecificatie configuratie van drie op elkaar afgestemde (software)componenten. Dit zijn de:

1. Privacy- en Verzend Module (PVM) die wordt gebruikt door de aanbieder (gele deel in onderstaand schema);
2. Centrale Module TTP (CMT) die wordt gebruikt door ZorgTTP (blauwe deel);
3. Doel- en Receive Module (DRM) die wordt gebruikt door SFK (rode deel).



Werking Privacy en Verzend Module (PVM)

Deze module wordt gebruikt door de aanbiedende partij; het gele deel in bovenstaand schema. De module kent een aantal functies. Allereerst wordt een aantal controles uitgevoerd op het aangeboden bestand. Daarna worden de persoonsgegevens omgezet in zogenaamde pre-pseudoniemen. Vervolgens wordt een scheiding aangebracht tussen de pseudoniemen (het sleuteldeel) en het bijbehorende datadeel (het SecureID). Beide delen worden vervolgens beveiligd met behulp van encryptie op zodanige wijze dat het sleuteldeel enkel kan worden geopend door ZorgTTP en het datadeel enkel kan worden geopend door de ontvangende partij, het doel.

Controle op aangeboden persoonsgegevens

Op de aangeboden persoonsgegevens worden logische controles uitgevoerd zoals:

'een datum moet voldoen aan het voorgeschreven formaat (dd-mm-eejj)'

Pre-pseudonimisatie

De eerste versleuteling die plaatsvindt bij de partij die beschikt over de te verzenden persoonsgegevens wordt ook wel pre-pseudonimisatie genoemd. Een voorbeeld van een pre-pseudoniem is de tekenreeks:

OS1C0039iaf4etutr0su85qv9gfsipex

In het voorbeeld vormen de eerste vier tekens (OS1C) de zogenaamde handtekening van het pseudoniem. Aan deze handtekening kan herkend worden dat het gaat om een *'Onbewerkte Sleutelwaarde' (OS)* van het *1e niveau (1)* voor het type pseudoniem *'C'.* Daarbij slaat het 1e niveau op de eerste bewerking bij de informatiebron en de 'C' op de gebruikte persoonsgegevens; Geboortedatum, Postcode en Geslacht. Het feitelijke pseudoniem wordt gevormd door de reeks van 28 tekens volgend op de handtekening.

De uitkomst van bovenstaand proces, in combinatie met de hieronder beschreven vervolgstappen, wordt verder in dit document in beeld gebracht onder de kop: 'voorbeeld werking pseudonimisatie'.

Centrale Module TTP (CMT) - Centrale pseudonimisatiesoftware:

De Centrale Module TTP ontvangt het in de PVM versleutelde bestand. Dit bestand bestaat uit twee onderdelen: een datadeel (het SecureID) en een sleuteldeel. Het sleuteldeel bevat de pre-pseudoniemen, deze worden door de centrale applicatie omgezet in de definitieve pseudoniemen. Hierbij wordt gebruik gemaakt van een specifieke SFK-sleutel. De centrale applicatie heeft geen toegang tot het datadeel. Alleen in de ontvangstapplicatie kunnen deze gegevens zinvol verder verwerkt worden. Na verwerking verstuurt CMT het bericht naar de ontvangende partij.

Doel- en Retour Module (DRM) - Lokale pseudonimisatiesoftware:

De ontvangstmodule wordt gebruikt door SFK. De DRM ontvangt berichten van de centrale TTP applicatie. Na ontvangst wordt de transportbeveiliging verwijderd en worden het datadeel en de pseudoniemen samengevoegd. Het samengevoegde bestand bevat gegevens die niet te herleiden zijn tot de oorspronkelijk aangeboden persoonsgegevens.

Het definitieve pseudoniem voor het eerder genoemde voorbeeld wordt:

FQ2C001rtd2wkt2rnm7wcdj5hyasbv8u

Aan de handtekening kan nu worden herkend dat het een pseudoniem bestemd voor SFK betreft (FQ) dat 2 maal bewerkt is en gebaseerd is op Geboortedatum, Postcode en Geslacht.

Bijlage 3: Voorbeeld werking pseudonimisatie SFK cohort

In onderstaande figuur wordt het in dit document beschreven proces in beeld gebracht aan de hand van de oorspronkelijk aangeboden data (input) en de aan het einde van het proces resterende data (output). Te zien is dat de waarde van de variabelen Uzovi en Verzekerdenummer niet gevuld is. De aanlevering van deze variabelen is namelijk optioneel (zie Bijlage 1).