Landelijk Coördinatiepunt Research Data Management

Omgang met pseudonimisering en sleutelbestanden bij kleinschalig onderzoek

Enkele basisstappen

VERSIE: DECEMBER 2019

LCRDM

Het Landelijk Coördinatiepunt Research Data Management is een landelijk netwerk van experts op het gebied van research data management (RDM). Het LCRDM maakt de koppeling tussen beleid en dagelijkse praktijk. Binnen het LCRDM werken experts samen om RDM-onderwerpen te agenderen die te groot zijn voor één instelling en die vragen om een gezamenlijke landelijke aanpak.

meer informatie: www.lcrdm.nl

COLOFON

Omgang met pseudonimisering en sleutelbestanden bij kleinschalig onderzoek
Enkele basisstappen

PUBLICATIEDATUM | december 2019
DOI | 10.5281/zenodo.3571046

LCRDM Taakgroep Pseudonimisering | **Simone van Kleef** (St. Antonius Ziekenhuis), **Jan Lucas van der Ploeg** (Universitair Medisch Centrum Groningen - UMCG), **Martiene Moester** (Leids Universitair Medisch Centrum - LUMC), **Henk van den Hoogen** (Universiteit Maastricht/liaison LCRDM adviesgroep), **Erik Jansen** (Universiteit Maastricht/liaison DataverseNL), **Tineke van der Meer** (Hogeschool Utrecht), **Francisco Romero Pastrana** (Rijksuniversiteit Groningen), **Jolien Scholten** (Vrije Universiteit), **Leander van der Spek** (Universiteit voor Humanistiek), **Ingeborg Verheul** (LCRDM)

Klankbordgroep | **Derk Arts** (Castor), **Marlon Domingus** (Erasmus Universiteit), **Laura Huis in 't Veld** (DANS), **Nicole Koster** (Universiteit Twente), **Karin van der Pal** (Leids Universitair Medisch Centrum – LUMC), **Alfons Schroten** (Universiteit Maastricht/MEMIC).

HANDOUT | Boudewijn van den Berg (LCRDM)
OPMAAK | Nina Noordzij, Collage, Grou
VERTALING | Gosse van der Leij

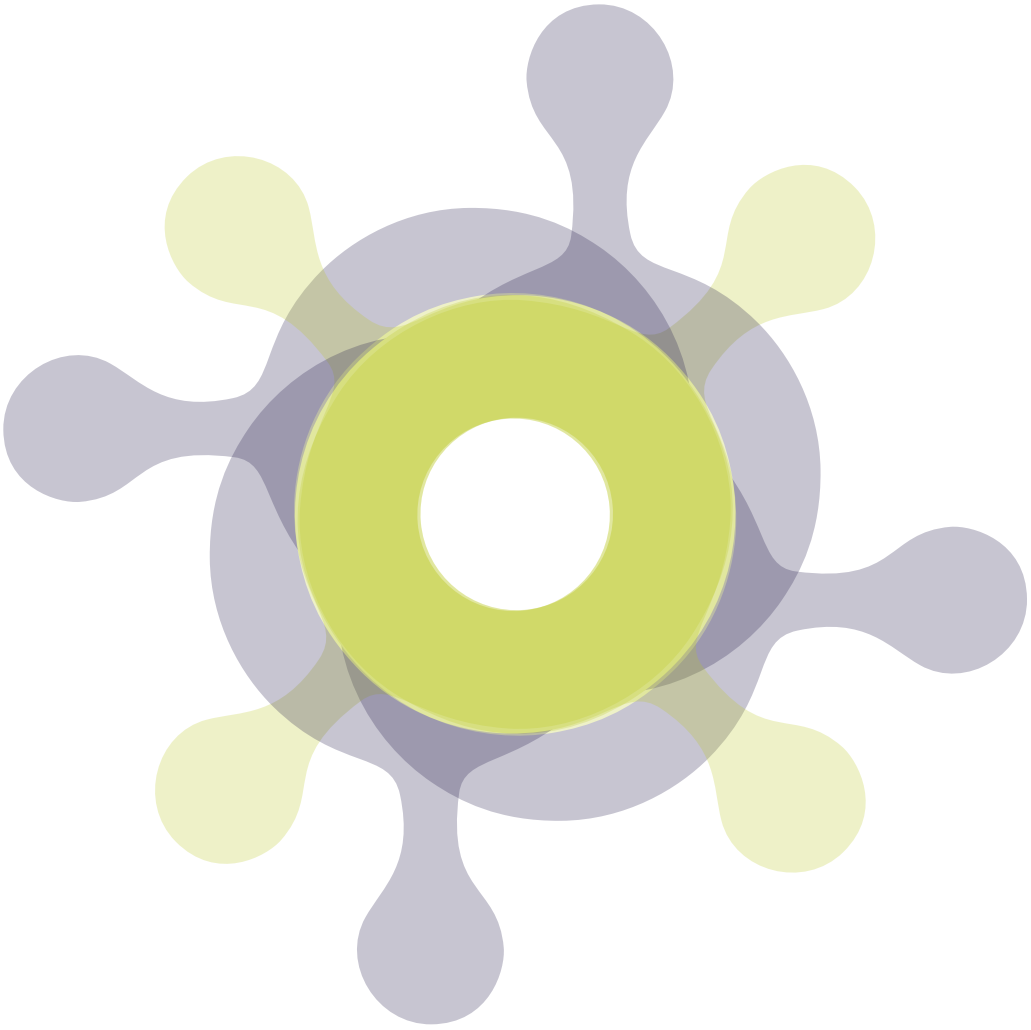
LCRDM

LCRDM wordt mogelijk gemaakt door

Omgang met pseudonimisering en sleutelbestanden bij kleinschalig onderzoek

Enkele basisstappen

Inhoud



Omgang met pseudonimisering en sleutelbestanden bij kleinschalig onderzoek

Enkele basisstappen

1] Inleiding

Iedereen die wetenschappelijk onderzoek doet met (bijzondere) persoonsgegevens heeft te maken met dezelfde vraagstukken: Hoe garanderen we de deelnemers aan het onderzoek dat hun persoonsgegevens veilig zijn opgeslagen? Hoe zorgen we ervoor dat de direct identificeerbare persoonsgegevens die nodig zijn voor de communicatie en organisatie van het onderzoek, gescheiden zijn van de onderzoeksgegevens?

Het is gebruikelijk dat datasets met (bijzondere) persoonsgegevens voor onderzoeksdoeleinden gepseudonimiseerd worden opgeslagen in een datamanagementsysteem. Bij grootschalige onderzoeken is er vaak budget om het pseudonimiseren te laten doen door Trusted Third Parties (TTP's), maar voor kleine onderzoeken is dat niet haalbaar. Daar is minder geld en tijd beschikbaar.

Wat is de werkwijze in zo'n geval? Met name het beheer van de sleutel(lijst) voor de koppeling tussen direct identificeerbare persoonsgegevens en de onderzoeksgegevens vraagt om zorgvuldigheid. Hoe bewaar je het? Wie mag erbij? Waar staat het? En hoe kun je ervoor zorgen dat de toegang tot de sleutel niet afhankelijk is van de kennis van één persoon, en ook voor de toekomst nog wel op de een of andere manier beschikbaar is? Bestaat er al een applicatie waarin dit allemaal goed geregeld is?

Een taakgroep van het LCRDM heeft in de periode februari – juni 2019 onderzocht of er praktische manieren bestaan om te pseudonimiseren bij kleinschalig onderzoek die ook relatief eenvoudig inzetbaar zijn bij andere instellingen. Mochten die niet gevonden worden, dan zouden er suggesties gedaan worden voor de eerste stappen die genomen kunnen worden bij het goed inzetten van pseudonimisering in kleinschalig onderzoek.

2] Wat is pseudonimiseren?

In een groot aantal vakgebieden is pseudonimisering al langer gangbaar. Het kan in onderzoek nodig zijn om onderzoeksdeelnemers te identificeren, bijvoorbeeld om de brondata te controleren of om personen over langere tijd te volgen. In zo'n geval kunnen de onderzoeksdata niet geanonimiseerd worden. Er wordt dan voor pseudonimiseren gekozen; niet alleen om de privacy van de onderzoeksdeelnemers te beschermen, maar ook uit oogpunt van wetenschappelijke integriteit (van onderzoek, onderzoeker en onderzoeksinstelling). De recente publiciteit rondom de invoering van de AVG heeft de belangstelling voor dit onderwerp extra aangewakkerd.[1]

Pseudonimiseren wordt hier opgevat als: het vervangen van de direct identificeerbare variabelen in een dataset door een pseudoniem. In sommige kringen wordt dit ook wel *coderen* genoemd.[2] Deze werkwijze hoeft er niet voor te zorgen dat de *gehele* dataset is gepseudonimiseerd. Als de dataset vrije tekstvelden bevat, staan daar potentieel nog direct herleidbare gegevens in. Daarnaast kan ook een combinatie van andere (niet direct identificeerbare) variabelen die van belang zijn voor het onderzoek, zorgen voor de identificatie van een individu.

De bedoeling van het pseudonimiseren in déze vorm is dan ook niet het verkrijgen van een anonieme dataset (of zo goed als anoniem[3]), maar om al tijdens de fase van dataverzameling de privacy van de onderzoeksdeelnemers te beschermen. Daarnaast moeten in het kader van wetenschappelijke integriteit de direct identificeerbare variabelen voor de onderzoeker achtergehouden worden.

1 Dat er veel behoefte was aan helderheid over dit onderwerp, bewees de tijd die het kostte om de taakgroep te vullen. Binnen 24 uur was er een tienkoppige taakgroep gevormd en een klankbordgroep van nog een aantal geïnteresseerden.

2 In bijlage 1 is een lijst van relevante definities opgenomen.

3 In tegenstelling tot gepseudonimiseerde data, is anonieme data op geen enkele manier te herleiden tot een persoon. Over omgaan met anonieme data is een andere LCRDM Taakgroep actief.

3] Wat is kleinschalig onderzoek?

De focus van de taakgroep lag bij de werkwijze van pseudonimiseren bij kleinschalig en kwantitatief onderzoek. Onder kleinschalig verstaan wij: onderzoek met een beperkte hoeveelheid participanten en/of weinig financiële middelen.

De beperking tot kwantitatieve data komt voort uit het feit dat het pseudonimiseren van kwalitatieve data (bijv video- en audiobestanden en transcripts hiervan) andere maatregelen vergt dan het simpelweg vervangen van direct identificeerbare gegevens uit een databestand met een code. Zo kunnen video's personen herkenbaar in beeld tonen, of iemand geeft tijdens een interview identificeerbare gegevens die vervolgens in het transcript terecht komen.

4] Aanpak

De taakgroep heeft allereerst op basis van eigen ervaringen en de beschikbare literatuur geïnventariseerd hoe pseudonimisering bij kleinschalig onderzoek wordt toegepast. Om een beter en vollediger beeld te krijgen van huidige pseudonimiseringspraktijken bij onderzoeksinstellingen is een enquête opgesteld. Deze is verspreid via verschillende kanalen, onder met name datamanagementondersteuners. Er is gevraagd naar de manieren waarop data worden gepseudonimiseerd en wie daarvoor verantwoordelijk is; of en welke software daarvoor wordt gebruikt; waar sleutelbestanden worden opgeslagen, wie er toegang toe heeft en wat ermee gebeurt wanneer een project afgerond is; of de instelling beleid heeft voor pseudonimisering en tegen welke problemen instellingen aanlopen. Van de 32 respondenten (zowel (enkele) onderzoekers als onderzoeksondersteuners) gebruikten 26 (een vorm van) pseudonimisering.

De belangrijkste conclusies die uit deze enquête getrokken kunnen worden, zijn:

1. De meeste instellingen gebruiken geen specifieke pseudonimiseringssoftware voor het pseudonimiseren van data. Sommige instellingen hebben hier wel tools voor, maar die kunnen niet rechtstreeks ingezet worden buiten het eigen onderzoek of de eigen instelling. Deze tools zijn op dit moment dus niet landelijk bruikbaar.
2. Bij de meeste instellingen ontbreekt beleid op het gebied van pseudonimisering of op een deelgebied ervan, bijvoorbeeld het omgaan met sleutelbestanden. De variatie in antwoorden laat ook zien dat de meningen over wat wel en niet mag, uiteenlopen.

In bijlage 2 is een uitgebreidere beschrijving van de resultaten opgenomen.

In aanvulling op de enquête is gekeken naar use cases in de dagelijkse praktijk van de instellingen van de taakgroepleden. Daarnaast is relevante wet- en regelgeving en andere gerelateerde documentatie bestudeerd (zie bijlage 3). Op grond hiervan is een lijst met basisstappen geformuleerd voor het pseudonimiseren van data.

5] Basisstappen

Dit rapport is bedoeld voor onderzoeksondersteuners, onderzoekers en/of onderzoeksinstellingen die nog weinig kennis van pseudonimisering hebben en die over onvoldoende tooling/infrastructuur beschikken.

Voorafgaand aan het doorlopen van onderstaande basisstappen, is het raadzaam eerst uit te zoeken of er bestaand beleid is met betrekking tot pseudonimisering in de instelling waar je werkt. Neem contact op met een specialist binnen je eigen organisatie als je vragen hebt over de implementatie van (één van) de hieronder beschreven maatregelen. Het instellingsbeleid prevaleert altijd boven de hieronder geformuleerde algemene basisstappen.

De taakgroep identificeert de volgende basisstappen die onderzoekers en onderzoeksondersteuners kunnen volgen bij het pseudonimiseren van datasets bij kleinschalig onderzoek.

1. Beschrijf in het datamanagementplan waarom en hoe je gaat pseudonimiseren, hoe de toegang tot het elders opgeslagen sleutelbestand en de dataset is geregeld en wat er met het sleutelbestand en de data gebeurt als het project is afgerond.

2. Identificeer de volgende categorieën in je data:
- Data die voor identificatie nodig zijn, om het onderzoek te organiseren of om te communiceren met de onderzoeksdeelnemers
  »»» Deze sla je op in het sleutelbestand
- Data die je nodig hebt voor analyse
  »»» Deze sla je bij voorkeur op in een datamanagementsysteem[4]
- Data die je niet nodig hebt (bijv. in geval van een aangeleverde dataset)
  »»» Deze data worden verwijderd.

3. Pseudonimiseer de data zo snel mogelijk, dus direct bij het verzamelen van de data. Als je een dataset met identificerende gegevens ontvangt van een andere partij, pseudonimiseer dan direct na ontvangst van de data.

4. Gebruik verschillende pseudoniemen voor verschillende datasets. Dit voorkomt dat data van deelnemers die in meerdere datasets voorkomen op pseudoniem gekoppeld kunnen worden.

5. Bewaar het sleutelbestand gescheiden van onderzoeksgegevens.

6. De toegang tot het sleutelbestand wordt bij voorkeur beheerd door iemand die niet betrokken is bij het onderzoek.

4 Een research datamanagementsysteem (DMS) is een programma waarmee je onderzoeksdata kunt opslaan en beheren. In een goed DMS worden alle handelingen in de onderzoeksdatabase vastgelegd (audit trail) en is de beveiliging adequaat geregeld.

7. Zorg voor adequate beveiliging en back-up van de data en het sleutelbestand.

8. Neem zowel technische als organisatorische maatregelen om te voorkomen dat ongeautoriseerde personen het sleutelbestand en de onderzoeksgegevens kunnen koppelen. Na afloop van de dataverzameling moet de toegang tot het sleutelbestand afgesloten worden voor de rol van onderzoeker.

9. Beperk de toegang tot het sleutelbestand, maar zorg ervoor dat er binnen de organisatie altijd wel iemand is met toegang tot het sleutelbestand.

6] Tot slot

Dit rapport beoogt niet het volledige spectrum te beschrijven waarbinnen pseudonimisering zou moeten worden toegepast. Er is sprake van een beperkte scope (pseudonimiseren voor kleinschalig onderzoek). Er is gebleken dat er een grote variatie is in wat de instellingen beschouwen als afdoende maatregelen voor pseudonimisering. Daarnaast is er onvoldoende eenduidigheid over wat pseudonimisering nu precies inhoudt.

Initieel had de taakgroep zich ten doel gesteld om op basis van een inventarisatie eisen en wensen te formuleren voor veilig beheer van de sleutelbestanden. De focus is uiteindelijk verschoven naar het formuleren van de basisstappen. De reden voor deze aanpassing is dat de inventarisatie geen duidelijke oplossingen/best practices opleverde, die voor andere instellingen direct toepasbaar zijn. De tools die in de antwoorden van de enquête zijn genoemd, waren vaak geen specifieke pseudonimiseringstools en zijn daarnaast meestal speciaal ontwikkeld voor grote onderzoeksinstellingen (UMCs/universiteiten). De inventarisatie gaf dus geen uitkomst waar onderzoekers van kleinschalig onderzoek direct mee aan de slag kunnen. De belangrijkste conclusie uit input die is opgehaald met de enquête was dat kennis van en ervaring met pseudonimiseren en goed beheer van sleutelbestanden nog bij veel onderzoekers en onderzoeksondersteuners tekortschiet. Met de basisstappen hebben we getracht enige randvoorwaarden aan te bieden voor een eerste start met pseudonimisering en het beheren van sleutelbestanden.

7] Aanbevelingen

Naast de basisstappen voor onderzoekers en onderzoeksondersteuners, doet de taakgroep tenslotte de volgende aanbevelingen:

1. Onderzoeksinstellingen hebben duidelijk en zichtbaar beleid nodig voor pseudonimisering en in het bijzonder voor het beheer van sleutelbestanden tijdens en na het onderzoek. Daarnaast moet er een infrastructuur zijn waar onderzoeksdata en identificerende gegevens gescheiden van elkaar opgeslagen kunnen worden, bij voorkeur in twee aparte adequaat beveiligde omgevingen.

2. Het is wenselijk dat LCRDM een netwerkdag organiseert met privacy-experts, beleidsmakers en onderzoeksondersteuners om eenduidige definities op te stellen voor privacy-gerelateerde concepten, met name voor de begrippen *pseudonimisering* en *anonimisering*. Er zijn geen duidelijke definities van privacy-gerelateerde concepten die breed gedragen worden in de onderzoekswereld. Afhankelijk van de context van data/onderzoek en de achtergrond van de deelnemers aan de discussie, kan dit leiden tot zéér uiteenlopende definities en onverenigbare standpunten. Voor de één is pseudonimiseren gelijk aan coderen, maar voor de ander is gecodeerde data niet per se gepseudonimiseerde data. Voor anderen hangt het af van de gebruikte technieken (encryptie, afgeschermde omgevingen). Voor weer een ander is gepseudonimiseerde data anoniem. Voor sommigen hangt anonimiteit af van de gebruiker van de data. De verschillende definities leiden niet alleen tot spraakverwarring, ook is zo niet altijd duidelijk of voor een onderzoek de juiste maatregelen getroffen worden. Het zou goed zijn om met onderzoeksondersteuners, beleidsmakers en privacy-experts tot breed gedragen definities en uitgangspunten te komen.

3. Een vervolgtaakgroep zou een uitgebreide inventarisatie kunnen doen van tools voor pseudonimiseren en het bewaren van sleutels, om op basis daarvan te komen tot requirements voor een algemeen beschikbare tool. Dit vraagt specifieke kennis van de taakgroepleden en een andere aanpak. Leden van de taakgroep moeten voldoende kennis van privacy en pseudonimisering hebben, zowel technisch als functioneel.[5]

5 Met de kennis van nu denken we dat een directe benadering van FG's of CISO's (corporate information security officers) meer kan opleveren dan een enquête onder ondersteunend personeel. Vaak hebben zij kennis van beleid en maatregelen rondom privacy en kunnen zo de weg wijzen naar de personen binnen de instellingen die zich bezighouden met pseudonimisering.

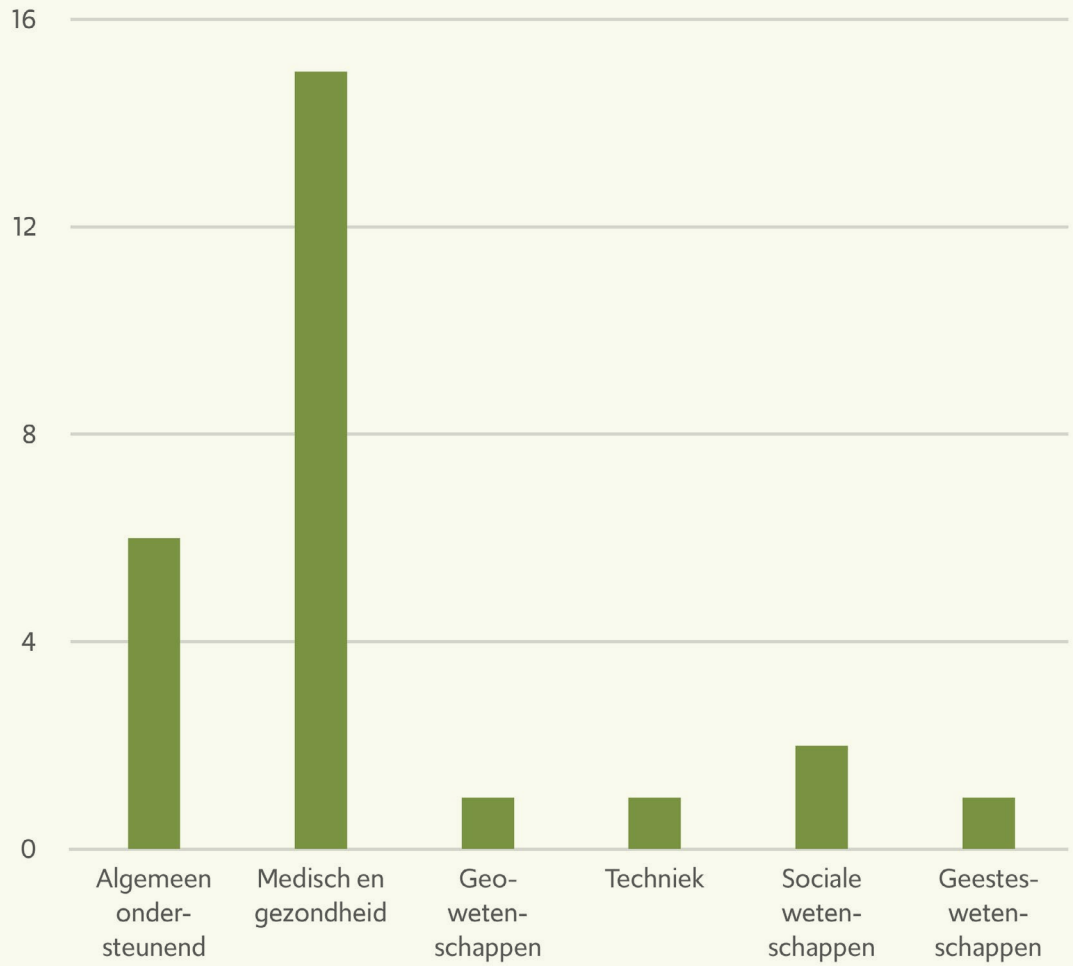
Bijlage 1 Definities

Persoonsgegevens	Alle informatie over een geïdentificeerde of identificeerbare natuurlijke persoon ('de betrokkene'); als identificeerbaar wordt beschouwd een natuurlijke persoon die direct of indirect kan worden geïdentificeerd, met name aan de hand van een identificator zoals een naam, een identificatienummer, locatiegegevens, een online identificator of van een of meer elementen die kenmerkend zijn voor de fysieke, fysiologische, genetische, psychische, economische, culturele of sociale identiteit van die natuurlijke persoon (AVG, artikel 4)
Pseudonimiseren volgens de AVG	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 (AVG, artikel 4). Bij pseudonimisering worden identificerende gegevens gescheiden van niet-identificerende gegevens en vervangen door kunstmatige identificatoren (Handleiding AVG, p. 27)
Pseudonimiseren volgens de definitie van de taakgroep	Het vervangen van direct identificeerbare gegevens door een pseudoniem. In medisch onderzoek wordt dit ook wel coderen genoemd. Deze definitie komt overeen met die uit de Handleiding AVG (zie hierboven), maar strikt genomen niet met de definitie in de AVG. Het vervangen van direct identificeerbare gegevens geeft geen garantie dat een specifieke betrokkene niet kan worden herleid. Andere gegevens in de dataset kunnen daar, al dan niet in combinatie met elkaar, ook voor zorgen.
Codelijst of sleutelbestand	Bestand met de combinatie van code/pseudoniem en bijbehorende direct identificerende gegevens.

Bijlage 2] Uitkomsten enquête

De achtergrond van de respondenten is vooral medische wetenschappen en ondersteuning van onderzoek. De redenen voor pseudonimisering en het type data dat gebruikt wordt voor onderzoek is divers, getuige de gegevens in de volgende grafieken.

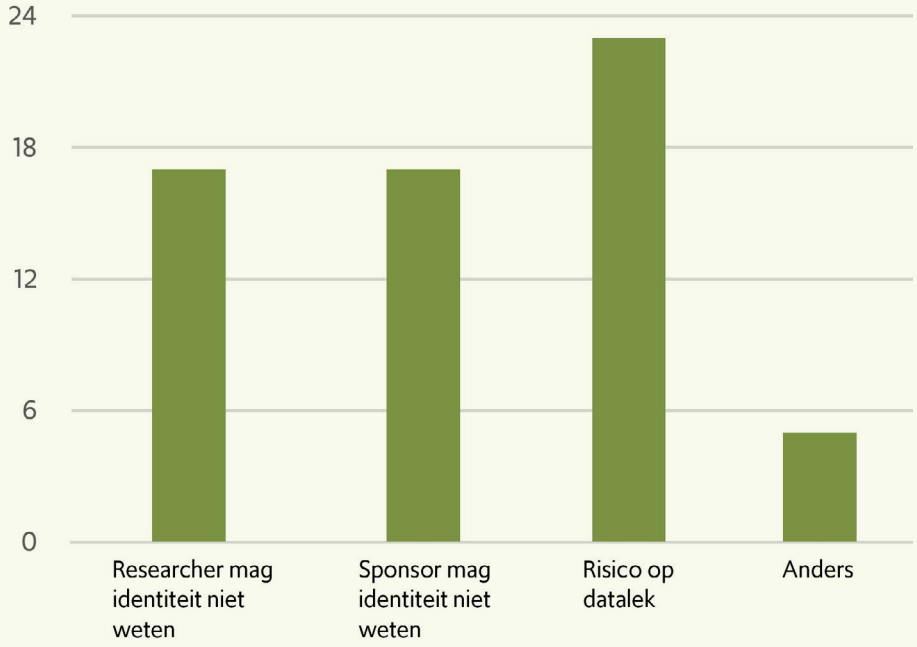
1] In welke onderzoeksdiscipline ben je werkzaam?



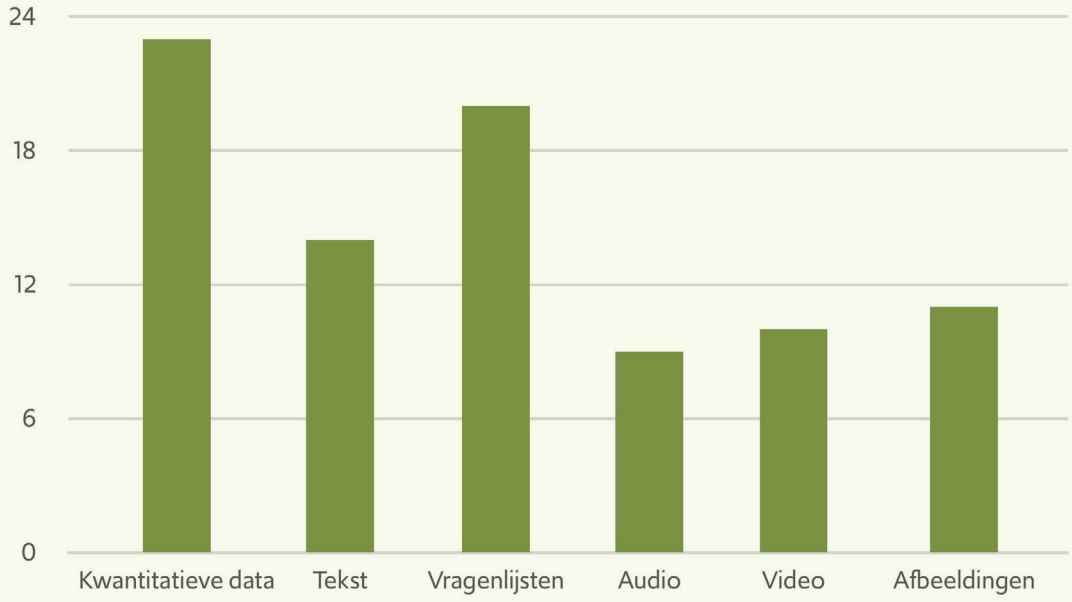
2] In welk type onderzoek wordt er gebruik gemaakt van gecodeerde persoonlijke gegevens? (bijvoorbeeld: kwalitatief/kwantitatief, WMO/nWMO, hoeveel proefpersonen)

De meeste respondenten antwoordden dat er in alle typen onderzoek wordt gepseudonimiseerd, met kwalitatieve en/of kwalitatieve data, WMO of nWMO en van twintig tot tienduizenden proefpersonen.

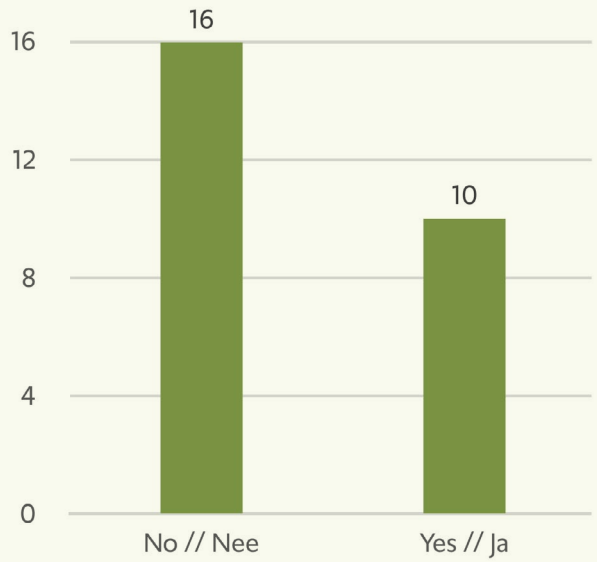
3] Waarom worden in jouw organisatie persoonlijke gegevens gecodeerd?



4] Welk type gegevens worden gecodeerd?



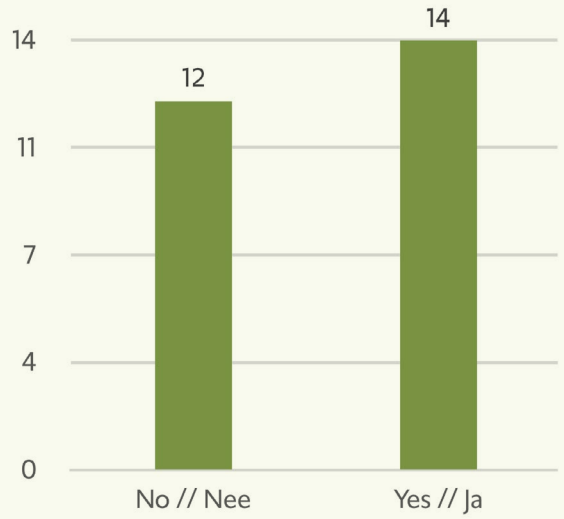
5] Heeft jouw organisatie beleid opgesteld hoe om te gaan met het coderen van persoonlijke gegevens?



6] Waar worden onderzoeksgegevens in jouw organisatie opgeslagen?

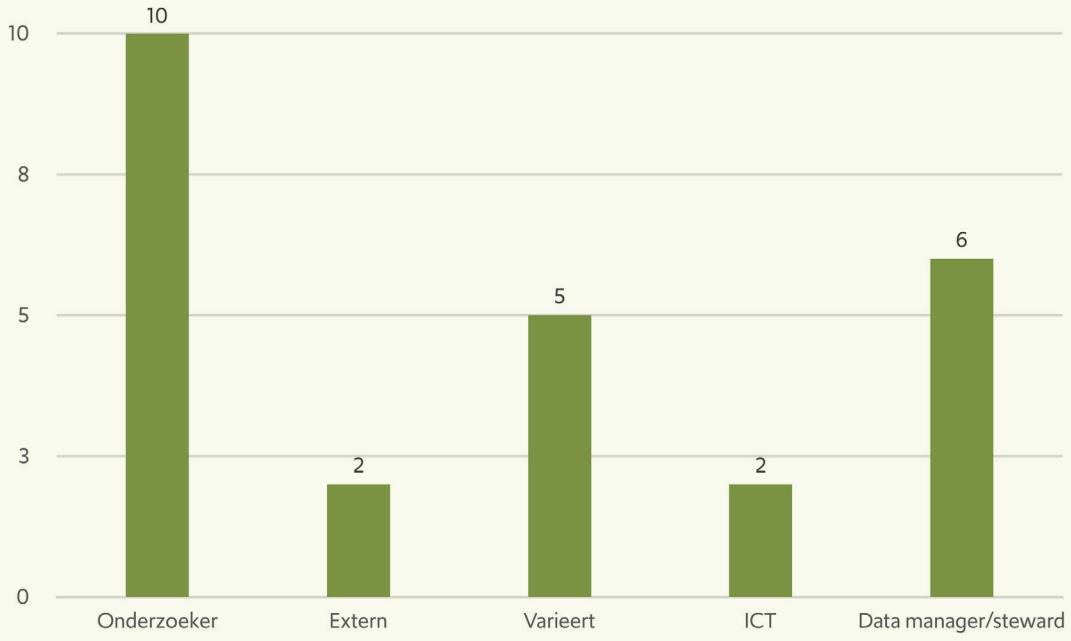
De meest genoemde locatie is de netwerkschijf: 75% van de respondenten geeft aan deze te gebruiken. Andere locaties die vaker genoemd worden, zijn: SURFdrive, eCRF/DMS en repositories. Een aantal respondenten noemt expliciet dat ze vermoeden dat onderzoekers ook privé data bewaren op persoonlijke schijven of in een dropbox.

7] Wordt er bij het coderen van data software of een ander technisch middel gebruikt?



Verreweg het vaakst wordt geantwoord dat een eigen onderzoeksplatform in deze functionaliteit voorziet. Andere opties die vaker worden genoemd zijn: SAS, encryptiesofware als VeraCrypt en een TTP service.

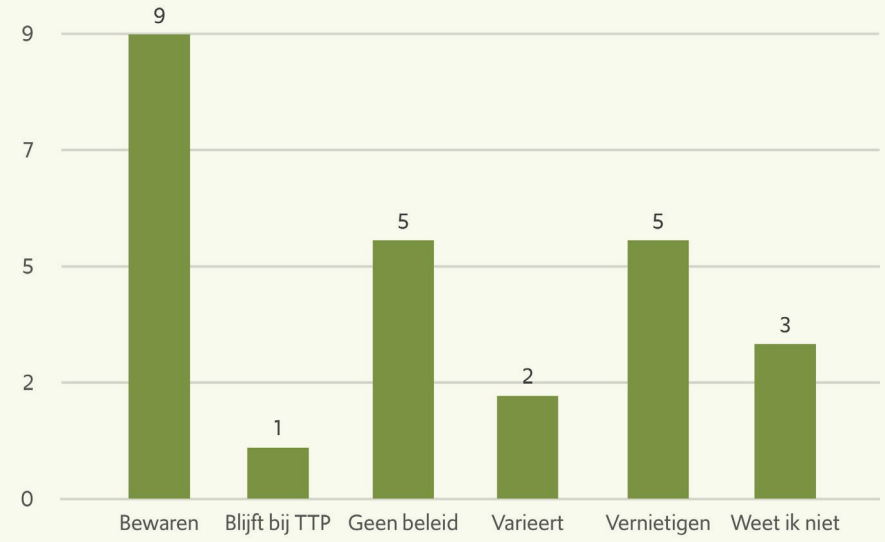
8] Welke functie heeft de persoon die verantwoordelijk is voor het coderen van de dataset?



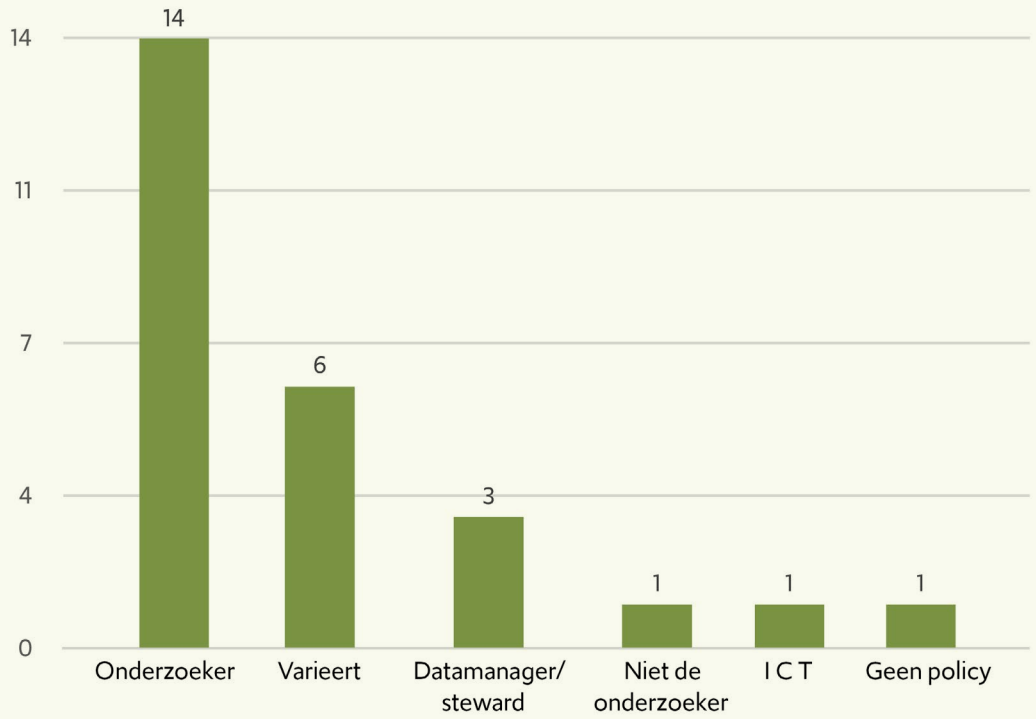
9] Hoe en waar wordt het sleutelbestand bewaard?

De antwoorden op deze vraag zijn zeer divers. Er wordt slechts vijf keer expliciet genoemd dat het sleutelbestand apart van de data wordt opgeslagen. Zes noemen als locatie dezelfde als die ze op de vraag over de opslag van de data noemden. Bijna iedereen geeft aan dat het bestand dan wel de map beveiligd wordt en alleen toegankelijk is voor geautoriseerde personen. Wie die personen zijn, daarover is geen overeenstemming. Een aantal geeft alleen de datamanagers toegang, maar in de meeste gevallen heeft het hele onderzoeksteam de beschikking over het sleutelbestand.

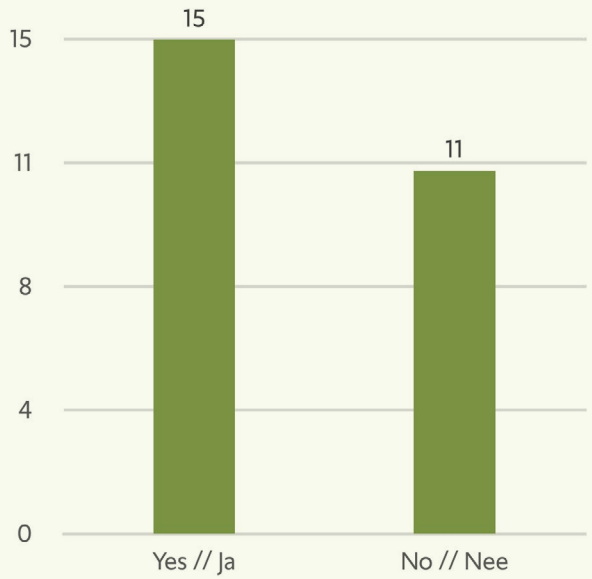
10] Wat gebeurt er met het sleutelbestand nadat het onderzoek is afgerond?



11] Wie heeft toegang tot het sleutelbestand?



12] Ondervind je problemen in dit proces?



Wanneer gevraagd wordt welke problemen men precies tegenkomt, is dit wat genoemd wordt:

– Onduidelijke richtlijnen: Wat is goed coderen? Hoe en waar bewaar je bestanden?
– Wat doe je met de bestanden na afloop van het onderzoek?
– Wie draagt de verantwoordelijkheid voor het sleutelbestand?
– Sleutelbestanden raken kwijt
– Sleutelbestanden zijn onvoldoende beveiligd
– Geen codeermogelijkheden voor audio en video
– Meerdere versies van sleutelbestanden
– Mensen re-identificeren participanten (of doen een poging daartoe)
– Er is geen monitoring en dus ook geen zicht op hoe onderzoekers met de data en sleutelbestanden omgaan.

Bijlage 3] Lijst van verwijzingen

– European General Data Protection Regulation (GDPR)
– Gedragscode gezondheidsonderzoek
– Infographic What is personal data?
– ISO 25237:2017 - Health informatics - Pseudonymization
– Wet medisch-wetenschappelijk onderzoek met mensen (WMO)
– Wet op de geneeskundige behandelingsovereenkomst (WGBO)
– Whitepaper on pseudonymization by the Data Protection Focus Group
– Pseudonimization guide for research data - Concept (F. Romero Pastrana, RUG)