# *Data-integriteit, data-opslag en privacy*

De wetenschap is de laatste paar jaren opgeschrikt door een aantal schandalen waarbij onderzoeks-gegevens zijn verzonnen of waarbij de correctheid van de gegevens niet meer te achterhalen is omdat onderliggende ruwe data “verdwenen” zou zijn. In het boek 'ontspoorde wetenschap' van Frank van Kolfschooten staan behalve de bij het brede publiek bekende zaken van Stapel, Smeesters etc.ook een aantal zaken waarbij het AMC een rol speelt. Wat opvalt is dat veel zaken onbevredigend eindigen. Daar zijn twee belangrijke factoren bij. Ten eerste zijn er vaak valide redenen waarom ruwe gegevens niet meer beschikbaar zouden zijn. “De privacy van de onderzochte personen mag niet geschonden worden”, “de harde schijf is gecrashed”, “ruwe data is bij een verhuizing verloren gegaan”. Allemaal dingen die in de dagelijkse praktijk van veel onderzoekers gebeuren, maar significant vaker optreden zodra buitenstaanders twijfelen aan de correctheid van gegevens. De tweede factor is dat onderzoekers nooit alleen onderzoek doen. Altijd zijn er collega's en organisaties bij betrokken die het wetenschappelijk wangedrag hadden moeten opmerken of medeauteur zijn. Zij hebben er dus allemaal belang bij dat de misdragingen worden gebagatelliseerd.
Het AMC heeft al sinds 2001 een ombudsman en een 'research code'. Dat is overigens een belangrijke reden dat het AMC vaker in het boek van van Kolfschooten optreed en veel andere ziekenhuizen niet. Wat in elk geval duidelijk is, is dat achteraf fraude vaststellen erg moeilijk is. Behalve achteraf onderzoeken is het dan ook noodzakelijk om fraude te voorkomen. Dat voorkomt ook ongewenste belangen van derden (collega's, bazen) die kunnen leiden tot het frustreren van het fraude-onderzoek.

## *Uitganspunten*

- De onderzoeker is verantwoordelijk voor de correctheid van de door hem of haar gebruikte data.
- Data-integriteit is een gedeelde verantwoordelijkheid van de onderzoeker en het onderzoeks instituut. Indien het instituut infrastructuur kan leveren waarin fraude bemoeilijkt wordt en dat niet doet is ze medeverantwoordelijk (en dus partij bij fraude-onderzoeken). Zeker als dat niet met significante investeringen gepaard gaat.
- Op elk moment (in het verleden, nu en de toekomst) moet vast te stellen zijn welke data er gebruikt is en op welke manier.
- Fraude moet zoveel mogelijk voorkomen worden, want achteraf bestraffen is bijna onmogelijk.
    - Zonder significant extra inspanningen van de onderzoekers
    - Zonder significante meerkosten voor de onderzoekers
- Ook klinisch onderzoek moet gebeuren met geanonimiseerde data.
- Dataintegriteit en kosten voor dataopslag zouden een vanzelfsprekend deel moeten zijn van een subsidie aanvraag. (maar zolang het dat niet is kan je daar ook niet met goed fatsoen achteraf een rekening voor sturen).

Dit stuk valt verder in drieen uiteen. Eerst een aantal bespiegelingen over hoe de voortgang van de techniek het mogelijk maakt om de problemen die ten grondslag liggen aan fraude grotendeels op te lossen. Daarna een voorstel hoe je dataopslag binnen het AMC (of landelijk) zou kunnen organiseren, zodat fraude in de vorm van datafabricage en malafide data selectie bijna uit te sluiten is. Plus een paar opmerkingen over data en andere zaken die wel in de loop van de tijd kunnen veranderen en over papier. Het derde deel gaat over de interactie tussen die onderzoeksomgeving en privacy.

# 1. Technische overwegingen

Eerst even een opmerking over de wet van Moore. De wet van Moore zegt dat ongeveer elke 18 maanden het aantal transistors op een chip verdubbelt. Dat is voor deze discussie niet direct van belang, maar er geldt een soortgelijke 'wet' voor data opslag. Elke anderhalf verdubbelt het aantal bytes dat je per geldeenheid kan kopen (in de praktijk gaat de verdubbeling voor opslag overigens nog sneller dan voor transistors). Dat wil zeggen dat als je data bewaart, dat je de tweede periode (van 18 maanden) nog maar de helft hoeft te betalen van wat het de eerste periode kostte. De derde periode is dat nog maar een kwart. Dit is een bekende reeks uit de wiskunde (1 + ½ + ¼ +... = 2) en de totale kosten voor eeuwige opslag zijn dus simpelweg 2 maal de kosten voor de eerste opslag. Kortom, als er genoeg geld is om bij de huidige prijzen 2 keer een  jaar opslag te betalen dan is er financiering tot het einde der tijden. Voor diegenen die wel eens met een archief te maken hebben is dit tegenintuitief, maar electronische opslag heeft dus fundamenteel andere eigenschappen dan 'fysieke' opslag. Dat verklaart ook mede waarom het onvermijdelijk is dat ziekenhuizen ernaar streven om het patienten dossier zo snel mogelijk volledig digitaal te hebben.
Er is natuurlijk ook een tegengestelde trend nl dat we steeds meer data verzamelen. Deze trend gaat echter langzamer, wat dus betekent dat uiteindelijk de totale dataopslag steeds goedkoper wordt.[1]
Deze overwegingen samen leidden tot een zeer belangrijke conclusie: digitaal opgeslagen informatie hoeft nooit meer weggegooid te worden wegens ruimtegebrek. Voor onderzoeksgegevens geldt nu meestal dat er een bewaartermijn is van 5 jaar. Dat is zinnig als de informatie op papier staat, maar voor digitale informatie is dat onzin. Vijf jaar is meer dan drie (twee keer 18 maanden) en dus kan (en moet) de informatie oneindig lang bewaard worden. Je kan het ook andersom formuleren: ***als je onderzoeks-gegevens wilt weggooien moet je daar wel heel erg zwaarwegende redenen voor aanvoeren.***[2]

Toch komt het nog geregeld voor dat data 'kwijtraken'. De redenen hiervoor zijn niet technisch, noch financieel. Het heeft te maken met de organisatie van het onderzoek. De onderzoekers zijn verantwoordelijk voor data-acquisitie, data-opslag én data-backup/archivering. Tot kort geleden was dat begrijpelijk, immers gegevens moesten snel toegankelijk zijn voor de onderzoekers en moesten dus op de lokale harde schijf staan; centrale opslag was te duur en te traag. Later werden USB en firewire zo snel dat het mogelijk werd om data op te slaan los van de eigen computer. Inmiddels zijn ook de netwerken zo snel geworden dat het niet meer strikt nodig is om data op de eigen computer te hebben. En als ze in de “cloud” staan weet je zelfs niet eens meer waar ter wereld ze fysiek staan. Dit heeft enorme consequenties voor hoe je binnen het onderzoek met je data kan omgaan. Het is n.l. niet meer nodig dat de onderzoeker zich bezig hoeft te houden met de hardware waar de data fysiek opgeslagen zijn. Daarmee ontstaat ook de mogelijkheid om data daar neer te zetten waar automatisch gearchiveerd wordt. Aangezien de meeste data-acquisitie systemen aan het netwerk gekoppeld zijn, zou ook de data-opslag automatisch kunnen worden geregeld. In dat geval is de onderzoeker alleen nog verantwoordelijk voor de metingen en gaat de rest “vanzelf”. Dat is niet alleen veel makkelijker voor de onderzoeker, maar dat voorkomt ook dat data niet meer traceerbaar zijn of, erger, dat gefingeerde data in een onderzoek terechtkomen. De boodschap van dit stuk is dan ook dat veel van de oorzaken van de problemen met data in verschillende schandalen van de laatste jaren simpel te voorkomen zijn door je dataopslag goed te organiseren en dat dat ook niet duurder hoeft te zijn dan het huidige systeem waarin fraude veel te makkelijk is.

---

1 Men heeft ook uitgerekend dat de totale beschikbare opslag*capaciteit* reeds een aantal jaren geleden de totale informatie*productie* van de hele mensheid gedurende haar hele historie is gepasseerd.

2 Dat geld ook voor e-mails in het kader van het onderzoek. Een onderzoeksinstituut waar nog e-mails moeten worden weggegooid omdat “de inbox vol is”, leeft nog in de 20e eeuw. Een voorzichtige schatting van het volume van alle e-mails in het AMC van alle medewerkers is dat dat waarschijnlijk minder is dan 1TB. Met een totale investering van het AMC voor die opslag van de orde van €100, af te schrijven over 4 jaar. En dan kost het nog maar €25 voor 4 jaar.

# 2. Naar een fraudevrije dataopslag

Als de individuele onderzoeker niet meer het beheer hoeft te doen van ruwe data, dan betekent dat automatisch dat opslag uniform geregeld moet worden. Aangezien er niet één logische manier is moet zo'n systeem verschillende manieren van toegang toestaan. Verder moet ook voor verschillende data vastliggen wie daar bij mag kunnen. Dat kan je low-level realiseren met een filesysteem of met een database of nog anders, maar gelukkig kan je dat tegenwoordig ook in het midden laten en gebruik maken van Uniform Resource Locators oftewel URL's. Hoe dat geimplementeerd wordt laat je dan in het midden en kan ook van URL tot URL verschillen. Voor de gebruiker ziet het er dan uit als een WWW of als een filesysteem.
Een manier om zoiets te implementeren is dat als persoon P op dag ddmmyyyy een meting doet op machine DevN van fabrikant F dat die data terug te vinden in de 'directory' `//amc.nl/data/devices/F/DevN/P/ddmmyyyy` (als de persoon zijn data wil terug vinden en voor de beheerder van het device) maar ook als b.v. `//amc.nl/projects/projectID/data/ddmmyyyy` (als je de data bij een project zoekt, en als je wel permissies hebt voor data in het project maar niet voor alle data van de machine en niet voor alle data van P) en als `//amc.nl/users/P/data/ExperimentN` (of welk pad persoon P ook maar prefereert, opdat de onderzoeker zelf de data in een voor haar logische manier kan organiseren).
Het spreekt vanzelf dat ruwe data, waar we het hierboven over hadden, altijd read-only wordt opgeslagen. Daarnaast kunnen (en zullen) er ook werkkopieën zijn[3]. Deze kunnen natuurlijk wel geschreven worden. In het ideale geval zal altijd zijn na te gaan hoe ze van de ruwe data afhangen. Soms is dat makkelijk en soms niet. Het meest praktische is als al deze informatie misschien (vooralsnog) wel lokaal verwerkt wordt, maar dat in elk geval op gezette tijden in een vergelijkbare datastructuur gesaved worden. De regel zou moeten zijn dat dat in elk geval moet als anderen bij die data moeten kunnen komen. Dat geldt dus is het bijzonder op het moment dat een artikel wordt ingestuurd. Op dat moment moet de onderliggende data “openbaar” en “gefixeerd” worden. Met openbaar wordt hier bedoelt dat duidelijk is welke data gebruikt is en dat een ander dan de onderzoeker zelf (bv de projectleider en de ombudsman) toegang heeft tot de data en de bewerkingen. Er zijn ook steeds meer geldschieters die eisen dat alle data ook echt openbaar wordt. Fixeren betekent hier dat die data van nu tot het einde der tijden toegankelijk zijn en niet meer veranderd kunnen worden.

## *Programmatuur/spreadsheets en databases*

Data op zich is meestal niet zo zinvol. Ze moeten ook bewerkt worden. In feite zit meestal de kennis in juist de analysemethode. Afhankelijk van het type onderzoeker en type data worden die analyses uitgevoerd met zelf geschreven programmatuur, in spreadsheets, middels sql-queries op databases of op nog andere wijzen. Het moge duidelijk zijn dat ook deze “gegevens” vastgelegd moeten worden. Ook dit moet gefixeerd kunnen worden op cruciale momenten in het project. Een of andere vorm van revisie controle is dus noodzakelijk. Op een wijze die uniform is voor het hele instituut, immers niet alleen de onderzoeker moet de versie die voor (bv) publicatie gebruikt is kunnen reproduceren, maar ook de projectleider. Dat betekend dan ook dat ook dát centraal geregeld moet worden. Liefst zo transparant mogelijk voor de gebruiker.

---

3 Ik probeer zelf altijd de ruwe data te inspecteren op eventuele herstelbare menselijke foutjes bij de acquisitie alsmede op overbodige informatie (kanalen die niet of verkeerd aangesloten waren, patient informatie,...), en dan een werkkopie te maken die schoon en gevalideerd is en in een format dat sneller en makkelijker te lezen is. Ik heb dus eigenlijk altijd een werkkopie en gebruik zelden de originele ruwe data.

### *Papieren archieven*

Zoals eerder terloops opgemerkt, digitale archieven zijn fundamenteel anders dan de papiere versies. Niet alleen zijn ze moeilijker toegankelijk, is een kopie maken lastig en de opslagkosten nemen niet snel met de tijd af, maar zijn constant of nemen zelfs toe. Digitaliseren lijkt een oplossing, maar is erg arbeidsintensief. Bovendien mag je daarna de originelen soms nog steeds niet weggooien. Voorbeelden zijn enquêtes, informed consent formulieren, tentamens en dergelijke. Papieren archieven zijn duur en veel moeilijker te controleren en beheersen dan digitale versies. In het belang van goed onderzoek lijkt het me dan ook noodzakelijk om te kijken naar alternatieven.

## 3. Patientenzorg en privacy

Behalve onderzoek is er natuurlijk ook patientenzorg in het AMC. In principe zijn alle gegevens die in dat kader gebruikt worden vertrouwelijk. Die vertrouwelijkheid kan op verschillende manieren on druk komen te staan. Ten eerste natuurlijk als login en paswoord in handen komen van een ander, maar totdat iemand een full-proof on-line biometrisch identificatie systeem ontwikkeld zal dat wel een probleem blijven. Een tweede manier is dat iemand niet geanonimiseerde gegevens per e-mail verstuurt of op een file exchange server zet waar ook derden toegang tot hebben (bv de systeembeheerder van het filesysteem waar de mail op staat. Gewone e-mail gaat als plain text over het netwerk en is daarmee ook onderschepbaar voor iedereen op een netwerk waar de mail langskomt). Vaak is men zich er niet van bewust dat men ongewild derden toegang kan hebben gegeven tot de data en helaas is het soms ook de enige praktische manier om gegevens uit te wisselen.
Manier drie is als er in een computer binnen het AMC ingebroken wordt. Daar zijn weer twee varianten van: een computer waar iemand op werkt met toegang tot het ZIS en een computer die voor andere dingen, zoals onderwijs of onderzoek gebruikt wordt. Aan de ene kant heeft inbraak in zo'n ZIS-computer grotere consequenties, maar ze zijn minder kwetsbaar omdat er maar een beperkte set programmatuur op kan worden gebruikt. Ook zijn ze niet van buiten toegankelijk. Inbraak op een onderzoeks computer is veel meer voor de hand liggend. Onderzoekers gebruiken veel verschillende programmatuur (en een diversiteit aan hardware), waarvan niet altijd de veiligheid te garanderen is. (Dat simpelweg verbieden kan overigens niet zonder in feite onderzoek onmogelijk te maken.) De gevaren voor wat betreft de privacy zijn bv dat (per ongeluk) niet alle gegevens op de computer anoniem zijn en vooral dat het als springplank kan dienen naar een, voorheen onzichtbare, machine binnen het netwerk die wel toegang heeft tot vertrouwelijke gegevens.

Het is daarmee wel duidelijk in welke richting de oplossingen moeten worden gezocht. Er moet een rigoreuze scheiding komen van privacy gevoelige data in de patientenzorg en de rest van de wereld. Als eerste stap moeten alle machines die niet met het ZIS hoeven te communiceren (én liefts ook die waarop “vreemde” software draait) van het AMC net af. Daar moet binnen het AMC een apart net voor komen. In feite is dat er al omdat we voor patienten en bezoekers (en langzamerhand ook voor steeds meer onderzoekers) een “public” net hebben. Dat hoeft niet strict een fysiek apart net te zijn zolang de machines uit public maar niet direct met de AMC-net computers kunnen communiceren en vice versa. In een later stadium moet dan de programmatuur waarmee de artsen communiceren met de buitenwereld (mail, browsers e.d.) ook van de AMC-net computers verwijderd worden. Dat zal wat meer tijd en aanpassing kosten, maar als straks iedere arts ook een smartphone of tablet heeft om informatie op te zoeken op het internet zal dat de normale patientenzorg niet in gevaar brengen. Communicatie van buiten het AMC-net met het ZIS moet dan gebeuren via een portal, zonder mogelijkheden om data direct naar een computer te downloaden.

Gaat dit dan niet het klinisch onderzoek frustreren? Absoluut niet. Zolang er maar een standaard manier komt om gegevens vanuit het AMC domein naar buiten te brengen. Essentieel is dat in die stap de data altijd geanonimiseerd wordt. De procedure wordt dan dat als een onderzoeker in het AMC gegevens van een patient wil hebben dat zij dan inlogt op een AMC portal en aan een bewaker van het fort AMC om die gegevens vraagt, die dan de gegevens opzoekt, anonimiseert en op een afgesproken plek in de bovenstaande research data structuur neerzet. Alleen op die manier kunnen we de privacy van de gegevens garanderen.

Terzijde: hiermee kunnen we ook de belangrijkste problemen met het EPD omzeilen. Stel dat een arts in de VU gegevens van het AMC wil hebben voor de behandeling van een patient die ook in het AMC bekend is. Beide ziekenhuizen hebben deze constructie van gescheiden patient- en onderzoeks-omgeving en beide hebben export én import routines voor klinische gegevens. De arts in de VU logt dan in op een portal van het AMC en, nadat zijn identiteit aldus is vastgesteld en eventueel na "overleggen" van een informed consent, vraag om de gegevens te exporteren. Hij krijgt een link terug waar de data staat voor incidentele consultaties, of hij spreekt van te voren waar dat zal zijn b.v. als het om een grotere groep gaat. Hij importeert die gegevens in de VU en deanonimiseert die met de patient gegevens die in de VU bekend zijn. Daarmee zijn de patienten gegevens over het netwerk van de boze buitenwereld verstuurd zonder dat op enig moment de privacy in gevaar is geweest. Een centrale opslag van gegevens hoeft hiermee niet en ook een gehackte computer van een huisarts of apotheker zal nooit verder komen dan een paar screenshots van een paar patienten.