PERSONAL HEALTH TRAIN

Het analyseren van gefragmenteerde data

Een voorwaarde voor een correcte schatting van een effect is dat alle relevante factoren zijn meegenomen in de analyse. Wanneer er een factor is weggelaten, zullen de schattingen van een effect vertekend zijn. Bijvoorbeeld in een analyse naar de effectiviteit van een medicijn, kan een vergeten factor ertoe leiden dat een medicijn veel effectiever lijkt dan werkelijk het geval is. Echter, als onderzoeker heb je niet altijd alle relevante factoren tot je beschikking. In zo'n situatie kan een onderzoeker ervoor kiezen om de data van bijvoorbeeld het Centraal Bureau voor de Statistiek (CBS) te koppelen aan de data van de onderzoeker om op deze manier meer relevante factoren op te nemen in de analyse. Om zo'n koppeling tot stand te brengen moet er eerst een aantal hobbels worden genomen. In dit artikel zullen de uitdagingen en oplossingen worden besproken.

Lianne Ippel & Johan van Soest

Over het algemeen spreken we van twee verschillende soorten van fragmentatie: horizontaal of verticaal. Horizontale fragmentatie betekent dat meerdere instellingen, bijvoorbeeld ziekenhuizen, dezelfde gegevens verzamelen over hun patiënten, zoals in tabel 1. Een analyse gebaseerd op deze gefragmenteerde data is eenvoudig omdat de analyseresultaten van elk ziekenhuis achteraf samengevoegd kunnen worden, zonder de privacy van de patiënten te schenden.

Zo eenvoudig is het echter niet wanneer we de gegevens van Ziekenhuis A willen koppelen aan gegevens van het CBS, zoals in tabel 2. In dit geval spreken we van verticaal gefragmenteerde data: Ziekenhuis A en het CBS hebben beide een ander deel van de gegevens van dezelfde mensen. Verticaal gefragmenteerde data zijn complex omdat we eerst moeten achterhalen welke gegevens van het CBS gekoppeld kunnen worden aan de gegevens van het ziekenhuis. Dit wordt nog extra bemoeilijkt omdat bij wet bepaald is dat de wetenschap geen gebruik mag maken van het Burgerservicenummer. Hierdoor is een unieke identificatie tussen de ziekenhuisgegevens en de CBS-gegevens onzeker. Daar komt bij dat het CBS gegevens bezit van veel meer individuen dan het ziekenhuis waardoor de kans van een foutieve koppeling groter is. Als we bijvoorbeeld geboortedatum gebruiken om de gegevens van een persoon van het ziekenhuis te koppelen aan de gegevens van het CBS, is de kans groot dat er meer mensen bekend zijn bij het CBS die op dezelfde datum geboren zijn. We hebben dan geen unieke koppeling en meer informatie is nodig om de juiste en unieke koppeling te vinden. Ten slotte staat de CBS-wet niet toe dat CBS-data uit hun faciliteiten gehaald worden, wat betekent dat de analyse altijd op de infrastructuur van het CBS gedaan zou moeten worden.

Het combineren van medische gegevens van een ziekenhuis met gegevens van het CBS kan dus alleen op een veilige manier gebeuren als de juiste juridische en ethische kaders aanwezig zijn die rekening houden met bijvoorbeeld de wet Algemene Verordening Gegevensbescherming (AVG), de CBS-wet, ethische aspecten van het onderzoek en daarnaast een technische infrastructuur die het mogelijk maakt om factoren van verschillende instellingen te combineren voor het onderzoek. Vervolgens moeten de data niet alleen juridisch maar ook praktisch herbruikbaar zijn voor onderzoek.

De FAIR-richtlijnen

In een notendop staat FAIR voor het bevorderen van het hergebruiken van data, voor de juiste doeleinden en onder de juiste voorwaarden. FAIR staat voor *Findable, Accessible, Interoperable* en *Reusable* (vindbaar, toegankelijk, voor meerdere doelen geschikt, en herbruikbaar) (Wilkinson et al., 2016).

Findable betekent dat alle digitale objecten een uniek identificatielabel krijgen waardoor er aan zo'n object gerefereerd kan worden. Een voorbeeld is DOI (Digital Object Identifier) een unieke code waarmee we digitale objecten kunnen vinden. De data moeten daarnaast een duidelijke beschrijving hebben zodat anderen ook begrijpen wat voor data ze gevonden hebben.

Accessible betekent dat er een protocol bestaat dat beschrijft hoe iemand toegang kan krijgen tot een digitaal object. Dit wil *niet* zeggen dat alle data al-

ZIEKENHUIS A					
Patiënt nr.	Leeftijd	Geslacht	Diabetes Type 2	Behandeling met diëtist	Roken
90123	45	M	Nee	Ja	Nee
90124	66	M	Ja	Ja	Ja
90125	51	V	Ja	Nee	Ja
90126	47	M	Nee	Nee	Ja
90127	63	V	Ja	Nee	Ja

ZIEKENHUIS B					
Patiënt nr.	Leeftijd	Geslacht	Diabetes Type 2	Behandeling met diëtist	Roken
298751	85	V	Nee	Nee	Ja
298752	71	M	Nee	Ja	Nee
298753	76	V	Ja	Nee	Nee
298754	41	M	Nee	Ja	Ja
298755	76	M	Ja	Nee	Ja
298756	54	V	Nee	Ja	Nee

Tabel 1. Horizontale fragmentatie van data: Ziekenhuis A en Ziekenhuis B hebben verschillende patiënten maar beide ziekenhuizen verzamelen dezelfde gegevens over hun patiënten

ZIEKENHUIS A					
Patiënt nr.	Leeftijd	Geslacht	Diabetes Type 2	Behandeling met diëtist	Roken
90123	45	M	Nee	Ja	Nee
90124	66	M	Ja	Ja	Ja
90125	51	V	Ja	Nee	Ja
90126	47	M	Nee	Nee	Ja
90127	63	V	Ja	Nee	Ja

CENTRAAL BUREAU VOOR DE STATISTIEK				
Id nr.	Leeftijd	Geslacht	Jaarinkomen euro	Eigenaar koopwoning
9020983	45	M	16.000-18.000	Nee
9020936	66	M	40.000-42.000	Ja
9027503	51	V	34.000-36.000	Nee
9024947	47	M		Ja
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
9054903	63	v	44.000-46.000	Ja

Tabel 2. Verticaal gefragmenteerde data: Ziekenhuis A en het CBS hebben gegevens van dezelfde mensen, maar ze verzamelen andere gegevens

tijd en voor iedereen beschikbaar zijn. De eigenaar van de data is te allen tijden verantwoordelijk dat de verzamelde data geen schade aan zullen richten aan degene van wie de data verzameld zijn. Daarom kan de eigenaar van de data beslissen om delen van de data af te schermen.

Interoperable betekent dat de data ook voor een ander doel gebruikt kunnen worden dan waar ze oorspronkelijk voor verzameld zijn. De data moeten daarom op zo'n manier gecodeerd worden dat de beschrijving van de data 'machine readable' is. Dit maakt het makkelijker om de computer een zoekopdracht te geven en alle relevante data te vinden. Daarnaast moeten de data opgeslagen zijn in een formaat wat te openen is door meerdere programma's zoals '.csv' in plaats van '.xlxs' omdat deze laatste alleen door het programma MS Excel te openen is.

Ten slotte *Reusable*, de herbruikbaarheid van de data, is de combinatie van de bovenstaande aspecten, met de juiste licentie zodat een andere onderzoeker weet wat toegestaan is om met de data te doen.

Data-stations, Treinen en Spoorwegen

Wanneer de juiste data gevonden zijn, en het ook juridisch en ethisch toegestaan is om de data te koppelen, kan het nog steeds voorkomen dat de data niet te downloaden zijn zoals bij CBS-data. Door middel van een nieuw initiatief, *Personal Health Train*, kunnen analyses naar verschillende data-partijen gestuurd worden zonder dat onderzoekers de data op hun eigen computer laden.

Het idee van de Personal Health Train is om in plaats van de data te verplaatsen naar de analyse, de analyse naar de data te brengen (Dutch tech center for life sciences, 2017). Dit heeft een aantal voordelen, bijvoorbeeld het beter beschermen van privacygevoelige informatie. Onderzoekers sturen namelijk alleen hun onderzoeksvraag en methode naar de data, en krijgen een antwoord terug, in plaats van dat zij de data van individuele personen inzien. Daarnaast is het verplaatsen van grote databestanden niet efficiënt: er ontstaan vele kopieën van hetzelfde databestand. Dit maakt het uitvoeren van de nieuwe AVG-wet nagenoeg onmogelijk. Als er echter maar één beschikbare versie van een databestand is, wat in een FAIR data-station staat, kunnen gegevens gemakkelijk aangevuld worden. Er zijn 'spoorwegen' die toegang kunnen verlenen tot deze stations. De spoorwegen zorgen ervoor dat onderzoekers alleen toegang tot de gegevens kunnen krijgen waar zij toestemming voor hebben. Wanneer iemand toegang heeft om de gegevens in het FAIR-data-station te analyseren kan de 'analysetrein' bij het station komen. Het station is daarmee ook verantwoordelijk dat alleen de toegestane informatie (resultaten) naar de onderzoeker worden verstuurd. Op deze manier kan de dataeigenaar zelfs na analyse toestemming voor toekomstig gebruik intrekken.

Applicatie

In het project FAIRHealth (Sun et al., 2018) – onderdeel van de route Verantwoordelijk Waardecreatie met Big Data (VWdata), een startimpulsprogramma van de Nationale Wetenschapsagenda – ontwikkelen we zowel technische als beleidsmatige oplossingen om gegevens afkomstig van De Maastricht Studie en het CBS in combinatie te analyseren. De Maastricht Studie verzamelt medische gegevens van participanten uit de regio Zuid-Limburg (Schram et al., 2014). We combineren deze medische gegevens met CBS-gegevens om de relatie tussen leefstijl, diabetes type 2 en zorgkosten te analyseren. Het combineren van deze gevoelige gegevens gebeurt in een veilige, afgeschermde, en privacybeschermende omgeving, waar niemand toegang toe heeft: de analyse wordt op de gecombineerde dataset geautomatiseerd uitgevoerd. Na de analyse wordt de gecombineerde dataset vernietigd en zullen alleen resultaten (bijvoorbeeld schattingen van de effecten) teruggestuurd worden naar de onderzoekers. Deze resultaten bevatten geen informatie die de identiteit van individuele patiënten kan onthullen. Zo waarborgen we de privacy van patiënten terwijl we toch een volledig

beeld verkrijgen van de leefomstandigheden van de patienten.

De technische en beleidsmatige oplossingen die ontwikkeld zijn in dit project zullen in de toekomst onderzoekers ondersteunen in het combineren van data; niet alleen van De Maastricht Studie en het CBS, maar ook data van andere partijen en in andere sectoren (bijvoorbeeld agricultuur of overheid). Hiervoor worden ook pilots uitgevoerd waarbij kennis uit dit project wordt hergebruikt, bijvoorbeeld met Vektis, de Nederlandse Zorgautoriteit en Zorginstituut Nederland.

Conclusie

Door databronnen met elkaar te koppelen kunnen we een beter totaalbeeld van de context krijgen. Om dat voor elkaar te krijgen en om zulke analyses op een verantwoordelijke en correcte manier mogelijk te maken, moeten juristen, IT'ers en methodologen samenwerken met onderzoekers. Verticale fragmentatie blijft dan ook een *hot topic* voor methodologen en computerwetenschappers, want naast de juiste infrastructuur om data te combineren zijn ook correcte analytische instrumenten nodig om tot valide uitkomsten te komen wanneer de data verticaal gefragmenteerd zijn.

Literatuur

Dutch Tech Center For Life Sciences (2017). *Manifesto of the Personal Health Train consortium.* <https://www.dtls.nl/wp-content/uploads/2017/12/PHT_Manifesto.pdf>.

Schram, M. T., Sep, S. J. S., van der Kallen, C. J., Dagnelie, P. C., Koster, A., Schaper, N., et al. (2014). The Maastricht Study: an extensive phenotyping study on determinants of type 2 diabetes, its complications and its comorbidities. *European Journal of Epidemiology, 29*, 439–451.

Sun, C., Ippel, L., Wouters, B., van Soest, J., Malic, A., Adekunle, O., et al. (2018). *Analyzing Partitioned FAIR Health Data Responsibly.* <http://arxiv.org/abs/1812.00991>.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*, 160018.

Lianne Ippel is postdoc bij het Institute of Data Science aan de Universiteit Maastricht.
E-mail: lianne.ippel@maastrichtuniversity.nl

Johan van Soest is postdoc bij het Institute of Data Science aan de Universiteit Maastricht en bij Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht UniversityMedical Centre+,
E-mail: johan.vansoest@maastrichtuniversity.nl

Henk Tijms

Sinterklaas en de laatste lootjes

Wie kent niet het Sinterklaaslootjesprobleem. Om de beurt trekt elk van n personen random een Sinterklaaslootje. De naam van elk van de personen staat op precies één lootje. Wat is de kans dat niemand zijn eigen lootje trekt? Een opmerkelijk en welbekend resultaat is dat bij niet al te kleine waardes van n de kans dat niemand zijn eigen lootje trekt praktisch gesproken niet van n afhangt en gelijk is aan $1/e \approx 0{,}368$. Deze benadering is in zeven of meer decimalen accuraat al vanaf $n = 10$. Wat wordt de oplossing van het probleem bij de volgende aanpassing van het probleem? Stel nu dat ieder die zijn eigen lootje trekt het lootje teruglegt en opnieuw random een lootje trekt totdat een ander lootje verkregen wordt. Wat is de kans dat de persoon die als laatste een lootje trekt met het eigen lootje blijft zitten? Deze kans is niet zo simpel te berekenen. Het antwoord is niet

$$\frac{n-1}{n} \times \frac{n-2}{n-1} \times \cdots \frac{1}{2} \times 1 = \frac{1}{n}.$$

De reden is dat deze berekening geen rekening houdt met het feit dat als iemand een lootje gaat trekken zijn eigen lootje al eerder getrokken kan zijn. Wel is $1/n$ een bovengrens op de gezochte kans. De sleutel tot de juiste oplossing is als volgt. Zonder beperking mag de aanname gemaakt worden dat elke keer door loting bepaald wordt wie als volgende een lootje trekt. De kans dat de laatste persoon blijft zitten met het eigen lootje verandert daardoor niet. De gemaakte aanname maakt het echter wel mogelijk een recursie op te stellen voor de berekening van de kans. Om de recursie op te stellen, is het inzichtelijk om met toestand (a,b) de situatie aan te geven dat a personen nog een lootje moeten trekken en dat voor b van deze a personen de lootjes met hun namen al getrokken zijn. In toestand (a,b) is de volgende persoon die een lootje gaat trekken met kans b/a een persoon wiens naam al eerder getrokken is en met kans $1 - b/a$ een persoon wiens naam nog niet getrokken is. Als in toestand (a,b) een persoon uit de groep van b personen wier namen al