Datascience in de psychiatrie

F.E. SCHEEPERS, V. MENGER, K. HAGOORT

ACHTERGROND De informatiesamenleving digitaliseert in een hoog tempo. Nieuwe technologie maakt reallife- en realtimeverzameling van tot nu toe ontoegankelijke informatiebronnen mogelijk. Dit creëert een bijna onmeetbare hoeveelheid dynamische data en daarmee mogelijkheden om in de psychiatrie tot nieuwe inzichten en verbeteringen in de behandeling te komen.

DOEL Verhelderen wat verstaan wordt onder big data en hoe een bigdata-aanpak de zorg kan veranderen in een constant lerend, door data gedreven en op de patiënt georiënteerd dynamisch systeem.

METHODE Korte beschrijving van een pilot in het UMC Utrecht waarbij het cross industry standard process for interactive data mining (CRISP-IDM) uitgevoerd werd en beschrijving van toepassingen in de toekomst.

RESULTATEN De beschreven aanpak en voorbeelden uit de literatuur laten zien dat er mogelijkheden zijn om snelle verbetering in de praktijk en nieuwe inzichten vanuit bestaande databronnen te realiseren.

CONCLUSIE De introductie van datascience in de psychiatrische praktijk biedt nieuwe mogelijkheden.

TIJDSCHRIFT VOOR PSYCHIATRIE 60(2018)3, 205-209

TREFWOORDEN big data, datamining, integratie

ARTIKEL

De wereld om ons heen is in een hoog tempo aan het veranderen. Digitalisering en nieuwe technologie zorgen ervoor dat vrijwel alles wat wij doen, meten en gebruiken vastgelegd wordt in digitale data. Dit is zeker ook het geval in de gezondheidszorg. Niet alleen ziekenhuizen, ggz-instellingen en huisartsen slaan immense hoeveelheden *care*- en *cure*data op via het elektronisch patiëntendossier, ook patiënten zelf verzamelen in toenemende mate gezondheidsdata via applicaties op hun smartphone of computer. Nieuwe sensoren die in *wearables* zitten zoals horloges, contactlenzen of digitale tatoeages gaan nog veel verder en kunnen 7 x 24 uur fysiologische of biochemische parameters meten zoals huidgeleiding, hartslagvariabiliteit en cortisolwaarde. Door deze metingen met GPS-tracking ook nog aan locatie te koppelen ontstaat een dynamische weergave van bepaalde waarden bij patiënten in hun natuurlijke habitat (Shoval e.a. 2017).

Ook in het (wetenschappelijk) onderzoek zorgen de beschreven ontwikkelingen voor een totaal andere kijk op gezondheidsmaten en -monitoring. Tot op heden werken we in epidemiologisch onderzoek of interventietrials met statische metingen, vaak in de spreekkamer van de dokter afgenomen en meestal op vooraf bepaalde (laagfrequente) momenten in de tijd. Uitkomsten zijn daarom niet altijd generaliseerbaar naar het echte leven. Sensordata kunnen het beeld van deze vertekende werkelijkheid nuanceren, maar vooral ook persoonlijker maken omdat patronen van individuele patiënten geanalyseerd kunnen worden. Hierdoor wordt zichtbaar hoe een specifieke patiënt reageert in interactie met de omgeving in de loop van de tijd (Wilhelm & Grossman 2010).

Ook de omgeving kan tegenwoordig steeds beter in beeld gebracht worden. In 2005 lanceerde Christopher Wild de term 'exposoom', een tegenhanger van het begrip 'genoom'. Het exposoom is de steeds veranderende omgeving, gemeten met sensoren die bijvoorbeeld luchtvervuiling, geluidsoverlast of het aantal mensen in de omgeving dynamisch kunnen vastleggen. Deze omgevingsfactoren kunnen begrijpelijkerwijs ook iemands mentale status beïnvloeden.

Tot slot zijn steeds meer dagelijkse activiteiten van mensen verbonden met 'de *cloud*'. Zoekgedrag via Google, sociaal gedrag via Facebook, Twitter en Instagram, maar ook huishoudelijke apparaten, de auto, de fiets, de mobiele

telefoon; alles is met alles gekoppeld, wat ook wel genoemd wordt *the internet of things* (IoT). Het IoT genereert een digitale weergave van ons dagelijks functioneren en kan dus gebruikt worden als bron van informatie die ook iets zegt over onze (mentale) gezondheid. Het IoT in combinatie met fysiologische of biologische sensordata van een individuele persoon wordt ook wel *the quantified self* (QS) genoemd.

Big data en gezondheidszorg

Als we spreken over big data gaat het bijna altijd over het grote volume van data. Big data betreffen echter meer dan dat. De 5 V's van big data (Marr 2015) staan voor *volume* (de grote hoeveelheid van data), *velocity* (de snelheid waarmee data verzameld en geanalyseerd kunnen worden), *variety* (de verschillende databronnen die aan elkaar gekoppeld kunnen worden), *veracity* (de 'echtheid' van data) en *value* (de waarde die je uit al deze bestaande data kunt genereren).

Als we ons beperken tot volume en snelheid zijn er in de gezondheidszorg genoeg voorbeelden van grote datasets die wereldwijd gekoppeld worden om met bigdata-analyses, met grote snelheid tot betere en meer valide inzichten te komen. In het genetisch (O'Donovan 2013) en beeldvormend (Thomson e.a. 2014) onderzoek bijvoorbeeld is het snel analyseren van grote databestanden met sterk verbeterde computerkracht dan ook niets nieuws. Echter, het blijven analyses van eenzijdige databronnen die verrijkt zijn met statische, zeer beperkte patiëntgegevens en nog summierdere omgevingsparameters. Terwijl in de psychiatrie deze nu juist zo van belang zijn.

Door koppelingen met nieuwe databronnen die het exposoom en de QS in beeld brengen, kunnen we komen tot variatie en echtheid. Immers, deze combinaties van verschillende databronnen benaderen steeds meer de realiteit van de (zieke) mens in interactie met zijn/haar omgeving. Dit leidt uiteindelijk tot waarde voor de gezondheidszorg. In het veld van bigdatastatistiek in de gezondheidszorg wordt gesproken over 'P4 medicine' (Hood & Galas 2008). Deze P's staan voor:

- *predictive* omdat je met complexe data-analyses veel beter kunt gaan voorspellen waardoor individuele mensen (psychisch) ziek worden, reageren op medicatie of een terugval krijgen;
- *preventive* omdat je met deze voorspellende waarde preventieve interventies kunt inzetten;
- *personalised* omdat vooral de QS individuele analyses mogelijk maakt;
- *participating* omdat QS-data van de patiënt zelf zijn en niet van de onderzoeker of arts. Hierdoor krijgt de patient een veel nadrukkelijkere rol in het diagnostiek- en behandelproces.

AUTEURS

FLOORTJE SCHEEPERS psychiater, medisch afdelingshoofd, afd. Psychiatrie, Hersencentrum, UMC Utrecht.

VINCENT MENGER, datascientist, promovendus, Information and Computing Sciences, Universiteit Utrecht.

KARIN HAGOORT, programmamanager Innovatie en big data, afd. Psychiatrie, Hersencentrum, UMC Utrecht.

CORRESPONDENTIEADRES

Floortje E. Scheepers, afd. Psychiatrie, divisie Hersenen, UMC Utrecht, Postbus 85.500, 3508 GA Utrecht.
E-mail: f.e.scheepers-2@umcutrecht.nl

Geen strijdige belangen meegedeeld.

Het artikel werd voor publicatie geaccepteerd op 31-5-2017.

Bigdata-aanpak in de praktijk

Overal in de wereld zijn organisaties bezig een datagedreven sturing op te zetten om processen en producten sneller en beter te maken. Het A&O fonds van de gemeenten in Nederland schreef onlangs een uitgebreid rapport over datagedreven sturing in gemeenten (A&O fonds 2017).

Veel van de beschrijvingen die in dat rapport staan, gaan ook op voor de gezondheidszorg. Het elektronisch patiëntendossier bijvoorbeeld kan als een bron van gestructureerde en ongestructureerde data veel beter benut worden voor directe zorgverbetering met bigdatastatistieken. Echter, ziekenhuizen hebben, los van het patiëntendossier, geselecteerde data (o.a. voor verplichte kwaliteitsmonitoring), meestal opgeslagen in datawarehouses. Doordat deze vaak in silo's zijn opgezet, missen ze flexibiliteit en vergen ze FAIR (*findable*, *accessible*, *interoperable*, *reusable*) opslag van gestructureerde data (Wilkinson e.a. 2016). Ook data uit onderzoeksprojecten zijn meestal in een gestructureerd datawarehouse opgeslagen.

Dynamische bigdata-analyses vragen juist om datapooling van zowel gestructureerde als ongestructureerde data, zoals teksten in de decursus in combinatie met medicatieoverzichten of labuitslagen. Vooral in de psychiatrie ligt veel informatie vast in grote hoeveelheden ongestructureerde teksten. Datapooling geeft veel meer mogelijkheden om verschillende bronnen flexibeler te koppelen en te analyseren, maar ook om losse punten in de datapool op verschillende manieren te benaderen in combinaties met andere data (een punt kan een andere betekenis krijgen door de datacontext te veranderen). Ook maakt een flexibelere dataopslag het gemakkelijker om met verschillende analyse- en visualisatietools te werken zoals SPSS, Excel, R

of Python en om geavanceerde analyses zoals *machine learning* en *text mining* toe te passen. Afhankelijk van de vraag, voorkeur van de datawetenschapper of noodzakelijke data kan deze kiezen voor de passendste tools.

Op basis van een datapool kunnen verschillende normprofielen voor patiënten gegenereerd worden (voor bijvoorbeeld specifieke kenmerken zoals er ook groeicurves en IQ-profielen zijn, als gaussverdeling). Voor de psychiatrie kun je denken aan de hoeveelheid (on)gezonde omgevingsfactoren (stressvolle life-events, urbanisatiegraad, aantal schoolwisselingen, enz.) in iemands leven die het risico op psychiatrische problemen positief of negatief kunnen beïnvloeden. De data van iedere nieuwe patiënt worden vervolgens gezien als een update van de datapool waarmee eerder gemaakte normprofielen aangepast kunnen worden aan de actualiteit. Immers, aanpassingen van een diagnose of behandeling van een patiënt zorgen soms voor verschuivingen in een individueel profiel en daarmee ook de normprofielen.

Beschrijvende analyse (wat gebeurt er, wanneer gebeurt het en waarom gebeurt het?) gaat zo over in voorspellende analyse (hoe profiteren we in de praktijk van deze kennis, welke profielen kunnen we genereren en wat kunnen we hiermee voorspellen?) naar ten slotte voorschrijvende analyse (welke beslissing nemen we op basis van het profiel en welke impact heeft die beslissing op al het andere?). TNO voorspelt dat onderzoek steeds meer datagedreven en minder hypothesegedreven zal worden (TNO 2013). Door complexe modelanalyses, waarbij een veelheid aan factoren en de meervoudige interactie tussen deze factoren onderling bepalend is voor de uitkomsten, verschuift het focus van causaliteit (oorzakelijke factor A leidt tot gevolg uitkomst B) naar risicoprofilering (de combinatie van factoren A, B en C bij persoon D leidt tot risico X op behandeleffect, terugval of symptoom) (Maathuis e.a. 2009).

Mogelijkheden voor toepassing

In de literatuur zijn verschillende voorbeelden van de toepassing van bigdata-analyses met klinische datasets beschreven, zowel in de algemene gezondheidszorg (Raghupati & Raghupati 2014) als in de psychiatrie (Kim e.a. 2015; Maenner e.a. 2016; Passos e.a. 2016). Voor de psychiatrie kun je denken aan risicoprofielen die de kans op een psychotische terugval kunnen voorspellen, de kans op bijwerkingen van specifieke medicatie of de kans om een depressie te krijgen.

Op de afdeling Psychiatrie van het UMC Utrecht werd in 2016 een pilotstudie verricht met data uit het patiëntendossier om te onderzoeken of een bigdata-aanpak implementeerbaar is en tot inzichten en verbeteringen voor de dagelijkse praktijk zou leiden. Daarbij werd geavanceerde visualisatiesoftware gebruikt die het mogelijk maakte dat zorgpersoneel uit de eigen praktijk de beschikbare data kon exploreren zonder specifieke technische kennis. Dit heeft twee grote voordelen: ten eerste kon men door deze aanpak exploratief en onbevooroordeeld zoeken naar nieuwe kennis en hypotheses. Daarnaast kon de dagelijkse praktijk direct actief meedoen in het proces.

Deze benadering had succes: met het *cross industry standard process for interactive data mining* (CRISP-IDM) vonden we 2 direct implementeerbare resultaten, en 29 interessante hypotheses die we nader kunnen onderzoeken (voor beschrijving van de methode zie Menger e.a. 2016).

Onder andere op het thema agressie wordt nu verder gezocht naar factoren die het risico verhogen of die voorspellend kunnen zijn. Denk hierbij aan de impact van de samenstelling van het verpleegkundige team op het optreden van agressie of het aantal gerapporteerde woorden in het dossier voordat een agressie-incident plaatsvindt. In de *proof-of-conceptfase* werd bijvoorbeeld zichtbaar dat er reeds twee dagen voor een agressie-incident een significante toename is van het aantal woorden in het dossier.

Ook het thema 'medicatie-effecten' wordt onder de loep genomen. Daarbij zoeken we naar individuele profielen die de kans op bijwerkingen van antipsychotica kunnen voorspellen.

Privacy en de-identificatie

Data van psychiatrisch patiënten zijn zeer persoonlijk en sensitief, daarom mag ook de ethische kant van big data niet onderbelicht blijven. Deels kan privacy worden gegarandeerd door patiëntdata te pseudonimiseren en identificerende variabelen te verwijderen uit de dataset (Fernandes e.a. 2013). Voor het de-identificeren van vrije tekst werd in het UMC Utrecht een eigen methode ontwikkeld, die zo veel mogelijk persoonsgegevens zoals persoonsnamen of adresgegevens automatisch verwijdert uit vrije tekst voor de punt graag invoegen: (Menger e.a. 2017). Daarbij moeten we opmerken dat wanneer een grote hoeveelheid variabelen over een patiënt bekend is, zelfs zulke maatregelen de kans op re-identificatie niet tot nul reduceren. Zodra dus gekoppeld wordt met externe databronnen zoals data verkregen door wearables is informed consent nodig van de patiënt.

Het koppelen van lokale data met data van andere zorginstellingen of bijvoorbeeld het CBS geeft een nog bredere blik op de patiënt en zijn of haar context. De implicaties hiervan voor de privacy van patiënten en in juridische zin dataeigenaarschap vragen echter nog grotere voorzichtigheid. Momenteel wordt gewerkt aan de implementatie van een dynamisch informed consent, waarbij de patiënt eigenaar blijft van zijn of haar data, en in de toekomst ook – met nadrukkelijke toestemming van de patiënt zelf – naar

databronnen buiten de eigen instelling gekeken kan worden (Williams e.a. 2015).

Conclusie

Omdat we in de psychiatrie vaak te maken hebben met complexe aandoeningen die fluctueren over de tijd en waarbij de interactie tussen kwetsbaarheid en omgeving een grote rol speelt, leent de psychiatrie zich bij uitstek voor een datagedreven lerende dynamische aanpak (Mcintosh e.a. 2016; Torous & Baker 2016). Immers, er zijn veel data (gestructureerd en ongestructureerd) in patiëntendossiers die amper benut worden en nieuwe technologie gaat nog veel meer relevante data genereren.

Deze aanpak vraagt om een nieuwe, flexibele wijze van datamanagement, een nieuwe invulling van privacy en data-eigenaarschap, maar de belangrijkste uitdaging is het creëren van een lerend klimaat waarin zorgprofessionals direct samenwerken met datawetenschappers binnen kort-cyclische *feedback loops* (*agile scrum*werkwijze) om de data directe klinische waarde te geven voor de praktijk. Analyse van de gehele, 'echte' patiëntenpopulatie met realtime- en reallifedata maakt het mogelijk de complexiteit in psychiatrische ziekten beter te begrijpen en om gericht op het individu te interveniëren.

LITERATUUR

- A&O fonds Gemeenten. Datagedreven sturing in gemeenten een verkenning van de veranderingen door het werken met big data. A&O fonds Gemeenten; 2017. (https://www.aeno.nl/wp-content/uploads/2017/03/AO_gemeenten_datagedreven_sturing_web_mrt17.pdf)
- Fernandes AC, Cloete D, Broadbent MT. Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records. BMC Med inform Decis mak 2013; 13: 71.
- Hood L, Galas D. P4 Medicine: Personalized, Predictive, Preventive, Participatory: a change of view that changes everything: A white paper prepared for the Computing Community Consortium committee of the Computing Research Association. 2008, http://cra.org/ccc/resources/ccc
- Kim JW, Sharma V, Ryan ND. Predicting methylphenidate response in ADHD using machine learning approaches. Int J Neuropharmacol 2015; 1-7.
- Maathuis MH, Kalisch M, Buhlmann P. Estimating high-dimensional intervention effects from observational data. Ann Statistics 2009; 37 (6A): 3133-64.
- Maenner MJ, Yeargin-Allsopp M, van Naarden Braun K, Christensen DL, Schieve LA. Development of a machine learning algorithm for the surveillance of autism spectrum disorder. Plos One 2016; 11(12): e0168224.
- Marr B. Big data: using SMART Big Data, analytics and metrics to make better decisions and improve performance. Hoboken: Wiley; 2015.
- McIntosh AM, Stewart R, John A, Smith DJ, Davis K, Sudlow C, e.a. Data science for mental health: a UK perspective on a global challenge. Lancet Psychiatry 2016; 3: 993-8.
- Menger V, Scheepers F, van Wijk LM, Spruit M. DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text. Telematics and Informatics 2017; http://dx.doi.org/10.1016/j.tele.2017.08.002.
- Menger V, Spruit MR, Hagoort K, Scheepers F. Transitioning to a data driven mental health practice: collaborative expert sessions for knowledge and hypothesis finding. Comp Math Methods Med 2016; 2016: 9089321.
- O'Donovan MC. What have we learned from the psychiatric genomics consortium. World Psychiatry 2013; 14: 291-3.
- Passos IC, Mwangi B, Kapczinski F. Big data analytics and machine learning: 2015 and beyond. Lancet Psychiatry 2016; 3: 13-5.
- Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst 2014; 2: 3. doi:10.1186/2047-2501-2-3.
- Shoval N, Schvimer J, Tamir M. Real-time measurement of tourists' objective and subjective emotions in time and space. J Travel Res 2017; doi: /10.1177/0047287517691155.
- Thomson PM, Stein JL, Medland SE. The ENIGMA consortium: large-scale collaborative analyses of neuroimaging and genetic data. Brain Imaging Behav 2014; 8: 153-82.
- Torous J, Baker JT. Why psychiatry needs data science and data science needs psychiatry: connecting with technology. JAMA Psychiatry 2016; 73: 3-4.
- Wevers C, Gijsbers G, red. TNO strategy&change. Innoveren voor gezondheid _ technologische en sociale vernieuwing in preventie en zorg-2.pdf 2013.
- Wilhelm FH, Grossman P. Emotions beyond the laboratory: theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment. Biol Psychol 2010; 84: 552-69.
- Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, e.a. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016; 3: 160018.
- Williams H, Spencer K, Sanders C. Dynamic consent: a possible solution to improve patient confidence and trust in how electronic patient records are used in medical research. Jmir Med Inf 2015; 3: e3.

SUMMARY

Data science in psychiatry

F.E. SCHEEPERS, V. MENGER, K. HAGOORT

BACKGROUND The information society is digitalising at a fast pace. New technology enables the collection of real life and real time information from sources that were inaccessible before. This creates an inordinate amount of dynamic data and, consequently, opportunities to introduce new insights and improvement of treatment in the field of psychiatry.

AIM To clarify the definition of big data and how a big data approach can reform care into a data driven, patient oriented dynamic system which is constantly learning.

METHOD Brief description of a pilot effected at the UMC Utrecht where the Cross Industry Standard Process for Interactive Data Mining (CRISP-IDM) was performed and description of applications in the future.

RESULTS The described approach and examples from literature show that there are possibilities to realise quick improvements in practice and implement new insights from existing data sources.

CONCLUSION Introduction of data science in psychiatric practice offers new prospects.

TIJDSCHRIFT VOOR PSYCHIATRIE 60(2018)3, 205-209

KEY WORDS big data, data mining, integration