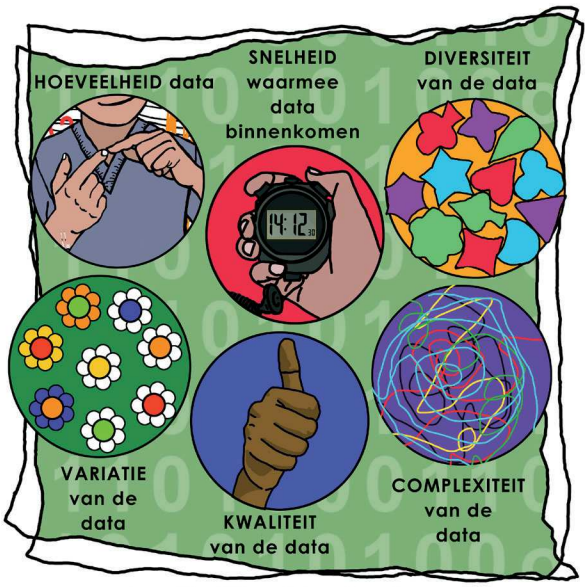

Definitie Big Data

Naast een gebrek aan consensus over de definitie, bestaan er nog andere inhoudelijke kwesties rondom Big Data waar zeker ook binnen het onderzoeksgebied van de epidemiologie, onduidelijkheid over is. Vragen als: *'Kunnen we met meer en onnauwkeurige data nog wel opzoek gaan naar causaliteit?', Vinden we onze voldoening straks in het rapporteren van nieuwe inzichten gebaseerd op patronen, trends en associaties op populatieniveau?'* Gaan we in de toekomst nog "old school" aan de slag met onze verschillende statistische software pakketten of opteren we daarvoor liever voor machine learning?', houden de gemoederen bezig. Kortom, ook in ons Epidemiologie veld is er onzekerheid over hoe we in de toekomst verder gaan met Big Data analyses. Vandaar dat dit in Epistel het onderwerp Big Data centraal staat, waarbij we vanuit diverse invalshoeken verschillende aspecten van Big Data, en de effecten van Big Data voor ons werkveld aan de orde zullen laten komen.

REFERENTIES:

1. Azmak O, Bayer H, Caplin A, et al. Using big data to understand the human condition: The kavli HUMAN project. Big Data. 2015;3(3):173-188.
2. Mathaiyan J, Chandrasekaran A, Davis S. Ethics of genomic research. Perspect Clin Res. 2013;4(1):100-104.
3. Lupton D. The commodification of patient opinion: The digital patient experience economy in the age of big data. Sociol Health Illn. 2014;36(6):856-869.
4. Costa FF. Big data in biomedicine. Drug Discov Today. 2014;19(4):433-440.
5. Boye N. Co-production of health enabled by next generation personal health systems. Stud Health Technol Inform. 2012;177:52-58 .
6. Mayer-Schönberger V, Cukier K. Big data: A revolution that will transform how we live, work, and think. Vol Great Brittain. John Murray An Hachette UK Company; 2013.
7. Boyd D, Crawford K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society. 2012;10:15(5):662-679.
8. Lazer D, Kennedy R, King G, Vespignani A. Big data. the parable of google flu: Traps in big data analysis. Science. 2014;343(6176):1203-1205.

Interview over Big Data prof. Robert de Jonge en dr. Mark Hoogendoorn

Door: Anouk Pijpe, Lilian Peters en Anouk Meijs

Mark Hoogendoorn: universitair hoofddocent kunstmatige intelligentie bij de VU

Robert de Jonge: hoogleraar/ afdelingshoofd klinische chemie in het AMC en VUmc

Wat is jullie relatie tot Big Data?

Mark Hoogendoorn: Mijn onderzoeksfocus ligt op machine learning in de gezondheidszorg. Hierbij werk ik veel samen met academische ziekenhuizen, in het VUmc onder andere met de IC en huisartsgeneeskunde. Daarnaast werk ik ook veel samen met de afdeling klinische psychologie bij de VU, waar ik heb meegewerkt aan het ontwikkelen van een app voor geautomatiseerde therapie bij depressie. Ik was met name betrokken bij het ontwikkelen van

de intelligentie in de app, zoals het bedenken van algoritmen en modellen om te voorspellen welke therapieën voor welke personen het meest geschikt zijn. Op dit moment is net een grote randomized trial in Europa afgerond die veelbelovend lijkt.

Robert de Jonge: In mijn eigen onderzoek kijk ik o.a. naar predictiemodellen voor therapie respons, zodat je de medicatie beter op de persoon kan afstemmen. De klinische chemie en ook de andere laboratoriumspecialismen doen nog te weinig met alle getallen die ze genereren. Bij een routine hematologie aanvraag kunnen wel 30 parameters gemeten worden, maar alleen de aangevraagde testen zoals de hemoglobine concentratie en het aantal witte bloedcellen wordt gerapporteerd. Achter die 30 parameters zitten nog veel meer ruwe data.. Daarnaast worden alle laboratoriumanalyses, denk daarbij ook aan die van de medische microbiologie, pathologie, klinische genetica en apotheek, vaak niet geïntegreerd geïnterpreteerd en gerapporteerd. Wij doen onderzoek of we met integrale analyse van alle ruwe en gerapporteerde data tot een betere diagnoses, prognose en therapie kunnen komen. Van labtest naar diagnose dus of van klinische chemie naar laboratoriumgeneeskunde, zoals ik dat in mijn oratie heb genoemd.

Wat is jullie definitie van big data?

Mark Hoogendoorn: Je hebt de 4 v's, maar ik vind big data sowieso een beetje een hype woord. Als wetenschappelijke term wordt het beperkter gebruikt. Die 4 v's zijn ook meer uit een consultancy achtige setting bedacht en dat klinkt natuurlijk lekker.

Robert de Jonge: Op het moment dat je andere tools nodig hebt dan de klassieke om je data te analyseren, dan noem je het big data. Dus als Excel niet meer werkt of als SPSS 20 uur aan het rekenen is dan zit je eigenlijk in de big data. Data science is de combinatie van big data en de analytics die je er op doet.

Denken jullie dat big data analyses door één bepaalde professional gedaan zouden kunnen worden, of is het alleen mogelijk met behulp van verschillende expertisen in samenwerkingsverband?

Mark Hoogendoorn: Ik zou zeker dat laatste zeggen. Ik denk dat als je kijkt vanuit de mensen die in de machine learning en kunstmatige intelligentie (KI) zitten, dat epidemiologen veel meer kennis hebben over hoe je een patiëntengroep selecteert. Wij zijn heel goed in de algoritmes, maar een bias die je introduceert door de keuze van data of van patiëntengroepen daar weten jullie veel meer van. Ook medisch inhoudelijk is jullie expertise heel waardevol. Een combinatie van meerdere expertisen bij elkaar is echt noodzakelijk. Ik denk dat de behoefte aan epidemiologen wel zal blijven om de uitkomsten van modellen goed te kunnen interpreteren, die zorgvuldigheid is heel belangrijk. Maar ook data scientists zijn belangrijk om de modellen op een goede manier toe te passen en niet gewoon maar te denken: ik gooi van alles in en tool, er komt iets uit en het zal wel goed zijn.

Wat zijn de verschillen en overeenkomsten in analyse technieken tussen epidemiologie en data science?

Mark Hoogendoorn: Ik denk dat de modellen vanuit Data Science in staat zijn om meer uit de data te halen. Om een voorbeeld te geven: wat een klassiek epidemiologisch model niet mee kan nemen is als er een bepaalde volgorde zit in een ontwikkeling, bv als een milde buikpijn steeds erger wordt. Natuurlijk is het feit dat het van mild naar erg gaat waarschijnlijk voorspellend voor een wat ernstiger ziekte dan als het van ernstig naar mild zou gaan. Het meenemen van die volgorde is iets wat je met machine learning goed kunt doen. Dat levert nieuwe inzichten op. Ook het omgaan met wat wij ongestructureerde data noemen (zoals medical images) is mogelijk met machine learning technieken, terwijl dat met de klassieke epidemiologisch methodes niet mogelijk is. Maar volgens mij gebruiken jullie ook een aantal technieken uit de machine learning, zoals een random-forest en decision trees. Het zijn misschien ook wel de termen, een beslisboom of logistische regressie noemen wij ook een machine learning techniek, maar dat is iets wat jullie heel vaak toepassen.

Verder is net als in de epidemiologie hetgeen wat het meeste tijd kost in het proces van machine learning het maken van features (variabelen) en het opschonen van de data, dat is zo'n 60% van het totale proces. Wat nu in opkomst is zijn de deep learning technieken, end-to-end learning, waarin je probeert dat stukje 'feature engineering' ook door de computer te laten doen. Maar op dit moment is het voor bepaalde types van data nog veelal handwerk, ook omdat je wilt begrijpen wat er gebeurt en je wil in overleg met een expert bepaalde zaken die mogelijke voorspellend zouden kunnen zijn vastleggen. Ik denk dat data scientists wel iets meer accepteren dat er wat ruis in de data zit. Bij ons zitten er vaak wel honderden variabelen in de dataset en die kun je natuurlijk niet allemaal in de meest optimale manier in de computer invoeren, dus je accepteert dat je meer variabelen en patiënten en ook meer ruis hebt.

Zijn de uitkomsten van de analyses wel te doorgronden en een goede basis voor bijvoorbeeld predictiemodellen: kortom is het nog uit te leggen?

Mark Hoogendoorn: Ik weet niet of de black-box achtige modellen altijd geschikt zijn voor het ziekenhuis. Ik denk dat we bij het analyseren van een scan op tumoren nog wel kunnen accepteren dat er sprake is van een black-box, waarbij er een cirkeltje getekend wordt waarin de tumor zit. Maar op het moment dat een arts naast een patiënt staat en advies krijgt over de dosis die hij zou moeten geven, dan moet hij wel weten wat er achter zo'n algoritme zit en waarom dat een goede dosis is voor deze patiënt. Ik denk ook dat de zorgvuldigheid waarmee je dingen in het medische domein moet implementeren veel meer testen vereist, zodat het op een goede manier geïmplementeerd wordt binnen zo'n organisatie.

Robert de Jonge: ons eerste artikel over een predictiemodel was heel lastig te publiceren. Er ligt toch nog een taboe op. Op congressen hoor ik dan dat we dat helemaal niet nodig hebben, maar de jonge mensen die zijn het al veel meer gewend. Een diagnose met een algoritme stellen, ze twijfelen daar niet eens meer over dat dat de toekomst is. Je kunt het haast niet afdwingen, het is een soort evolutie die plaats vindt, de jonge generatie die omarmt dat veel makkelijker.

Worden machine learning modellen al veel toegepast in het ziekenhuis?

Mark Hoogendoorn: Er loopt op dit moment een trial met een algoritme in het elektronische patiëntendossier op de intensive care van de VUmc die een advies geeft van wat voor dosering je van een bepaald type medicatie zou kunnen geven om de optimale stabiele hoeveelheid werkzame stof in het bloed te houden, en daar zie je dat de artsen daar wel degelijk naar kijken. Dit project wordt geleid door intensivist Paul Elbers. Dat is het eerste voorbeeld waar ik bekend mee ben van een model dat ook daadwerkelijk gebruikt wordt bij het VUmc. De IC is natuurlijk ook een heel mooi voorbeeld omdat het zo'n data gedreven vakgebied is en ze zijn gewend naar al die data te kijken en te proberen dan de behandeling te optimaliseren. Dus ik denk dat dat ook echt het eerste vakgebied is wat zo snel zo kan denken.

Robert de Jonge: In het ziekenhuis worden commercieel verkrijgbare testen als medische hulpmiddelen beschouwd en moeten aan wettelijke eisen voldoen voordat ze op de Europese markt mogen worden toegelaten. Maar als je een predictiemodel maakt op basis van deze testen, hoe werkt dat dan? Als een model ook de diagnose gaat voorspellen, dan kom je in een heel interessant domein terecht. We zitten dus ook wel deels gevangen in die regelgeving, geneeskunde is echt een low risk environment. Je wil echt wel zeker weten dat wat je doet ook juist is, dat maakt dit soort ontwikkelingen soms ook best wel lastig.

Zijn er opvallende verschillen tussen de academische wereld en het bedrijfsleven?

Mark Hoogendoorn: In het ziekenhuis ben je soms 5 jaar verder voordat je überhaupt resultaten van een onderzoek hebt. Als je dan kijkt naar de software ontwikkelingen en de KI algoritmes die allemaal verbeteren dan is 5 jaar een eeuw. Dat probleem hebben we nu met een app die we hebben ontwikkeld. Als je naar de app kijkt, het is gegeven de huidige ontwikkelen niet een ontzettende fancy app, maar dat is logisch want we zijn 10 jaar geleden begonnen met het ontwikkelen en nu is het eindelijk tijd voor de RCT. Bedrijven hebben wel de vrijheid om een app te ontwikkelen en die direct op de markt te zetten, de snelheid waarmee die ontwikkelingen gaan is enorm. Natuurlijk is het bewijzen van effectiviteit op een wetenschappelijk manier wel weer heel belangrijk. Dit maakt het lastig.

Robert de Jonge: Ik denk dat dat die hele integratie van de data nu wel echt exponentieel het vakgebied veranderen. Ik zie zoveel om me heen, ook bij bedrijven, dat ik me ook wel een beetje zorgen maak. Waar staan wij nog straks als wij daar niet ook vol in gaan ontwikkelen. Nu zijn we nog deels beschermd door de muren van het ziekenhuis, bedrijven kunnen niet bij het EPD en de data, maar als patiënten straks dingen gaan delen verandert dat misschien. Wat als mensen hun eigen DNA of bloed onderzoeken en dat delen met een app dat met een zelflerend algoritme een diagnose stelt? Het past allemaal bij de trend dat kennis straks voor iedereen toegankelijk is en dat de patiënt achter het stuur komt te zitten.

Daarom is die connectie met value based healtcare ook heel interessant, daarin gaat het om het creëren van meerwaarde (betere uitkomsten op voor de patiënt relevante maten ten opzichte van de kosten om deze uitkomsten te realiseren). Met machine learning en KI ontwikkelde algoritmen kunnen tot meerwaarde in de zorg leiden. Vanuit het laboratorium bezien gaat het steeds minder om wie de data genereert, maar wel wat je daar dan vervolgens mee doet.

Welke samenwerkingsverbanden lopen er op het gebied van big data?

Mark Hoogendoorn: Het hele proces van het verkrijgen van data voor mijn analyses binnen het VUmc kost veel tijd kost en het is met name gebaseerd op goed contact met specifieke personen binnen het ziekenhuis. Daarom zijn we op initiatief van de eerdergenoemde intensivist Paul Elbers bij het AMC/VUmc begonnen met een netwerk van personen uit verschillende disciplines die met big data werken: Amsterdam Medical Data Science. Hiermee willen we zorgen dat mensen met elkaar gaan praten en niet iedereen in zijn eigen silo'tje werkt en z'n eigen ding doet en niet van elkaar geleerd kan worden. Ook willen we dat de data scientists met de artsen samen gaan zitten zodat de artsen iets meer begrijpen wat de data scientists doen, en andersom natuurlijk ook. Er zijn inmiddels al meer dan 400 mensen lid van het netwerk.

Vanuit de VU hebben we ook ACBA (Amsterdam Centre for Business Analytics), geleid door onder andere Frans Feldberg en Ger Koole. Het is een instituut wat kennis samenbrengt op het gebied van big data en data science. Dat is inclusief de informatica en wiskunde kant, maar ook de business kant. Als je iets met big data wilt gaan doen, wat betekent dat dan voor je organisatie? Hoe kun je dat het beste implementeren? Misschien wordt ook je business model wel anders, wordt je bv leverancier van predictieve modellen. Hier denken we over na binnen het instituut. Het is zowel gericht op onderzoek als onderwijs.

Robert de Jonge: ACBA organiseert ook een een post-graduate opleiding voor bedrijven. In het geneeskunde domein willen we ook graag zo'n opleiding, maar dit toevoegen aan de reguliere opleiding was tot op heden nog een brug te ver. Binnen onze beroepsvereniging maar ook binnen de geneeskunde was er behoefte aan een cursus, dus hebben we besloten om de opleiding van een jaar die al bestond nu als pilot in 7 dagen te stoppen en eens te kijken hoe dat zou gaan. Die 7 dagen geven een goed beeld van wat data science nu inhoudt en wat zou het kunnen betekenen in de zorg en onderzoek, maar je bent uiteraard niet meteen een data science expert na deze 7 dagen. De vraag is dus hoe we daar in de toekomst mee om willen gaan. In het onderzoek zijn steeds meer mensen bezig met data science merken we. Dit was ook de reden waarom we met een groot aantal mensen uit het Amsterdam UMC en de VU/UVA Amsterdam Medical data Science hebben opgericht. De implementatie in de zorg gaat uiteraard veel langzamer.

Waar staan we over 10 jaar op het gebied van big data?

Mark Hoogendoorn: Dat is wel een goede vraag. De ontwikkelingen aan de technische kant gaan echt enorm snel. Maar als je ziet hoe lang het duurt voordat iets daadwerkelijk in het ziekenhuis is ingevoerd, dan gaat daar wel echt heel veel tijd overheen. De ontwikkelingen in de techniek zullen de komende 10 jaar ook nog wel zo door gaan. Als je ziet wat er de afgelopen 5 jaar allemaal gebeurd is binnen de KI in machine learning dan zijn dat enorm snelle ontwikkelingen geweest.

Ook binnen het onderzoek komt de focus steeds meer te liggen op wat het nu inhoudt wat er geleerd is. Ik zei eerder dat black-box methoden iets is wat je misschien niet altijd wilt gebruiken in de medische context, maar je ziet dat nu steeds meer onderzoek zicht richt op het begrijpen wat er intern in zo'n netwerk plaats vindt, een soort visualisatie. Dat is denk ik ook wel een trend die over de komende jaren zal doorzetten. Zo'n netwerk heeft een enorme rijkheid van uitdrukkingskracht, het begint met de vraag wat zou er voorspellend zijn van die ruwe data, naar een daadwerkelijke voorspelling. Nu is het nog veel black-box, maar ik denk dat het dus wel minder black-box zal worden, zodat je wat beter kunt begrijpen wat er plaats vindt daar binnen in. Dat is denk ik een ontwikkeling die hopelijk ook in de komende jaren nog doorzet.