

Leo Ottes

BIG DATA IN DE ZORG

Big Data in de zorg

De serie 'Working Papers' omvat studies die in het kader van de werkzaamheden van de WRR tot stand zijn gekomen. De verantwoordelijkheid voor de inhoud en de ingenomen standpunten berust bij de auteurs. Een overzicht van alle webpublicaties is te vinden op www.wrr.nl.

Wetenschappelijke Raad voor het Regeringsbeleid
Buitenhof 34
Postbus 20004
2500 EA Den Haag
Telefoon 070-356 46 00
E-mail info@wrr.nl
Website www.wrr.nl

Big Data in de zorg

Leo Ottes

Rapporten aan de Regering nrs. 68 t/m 95 zijn verkrijgbaar in de boekhandel of via Amsterdam University Press (www.aup.nl).
Alle *Rapporten aan de Regering* en publicaties in de reeksen *Verkenningen* en *Working papers* zijn beschikbaar via www.wrr.nl.

Vormgeving binnenwerk: Textcetera, Den Haag
Omslagafbeelding: Textcetera, Den Haag
Working Paper nummer 19

ISBN 978-94-90186-28-9

WRR, Den Haag 2016

Alle rechten voorbehouden. Niets uit deze uitgave mag worden verveelvoudigd, opgeslagen in een geautomatiseerd gegevensbestand, of openbaar gemaakt, in enige vorm of op enige wijze, hetzij elektronisch, mechanisch, door fotokopieën, opnamen of enige andere manier, zonder voorafgaande schriftelijke toestemming van de uitgever.

Voor zover het maken van kopieën uit deze uitgave is toegestaan op grond van artikel 16B Auteurswet 1912 j^o het Besluit van 20 juni 1974, Stb. 351, zoals gewijzigd bij het Besluit van 23 augustus 1985, Stb. 471 en artikel 17 Auteurswet 1912, dient men de daarvoor wettelijk verschuldigde vergoedingen te voldoen aan de Stichting Reprorecht (Postbus 3051, 2130 KB Hoofddorp). Voor het overnemen van gedeelte(n) uit deze uitgave in bloemlezingen, readers en andere compilatiewerken (artikel 16 Auteurswet 1912) dient men zich tot de uitgever te wenden.

INHOUD

Voorwoord	7
Samenvatting	9
1 Inleiding	11
1.1 De zorg genereert veel gegevens	11
1.2 Kennis van oorzaken van ziekten en aandoeningen is beperkt	11
1.3 Kwaliteit en doelmatigheid van de zorg	12
1.4 Big Data to the rescue?	12
1.5 Leeswijzer	13
2 Wat is Big Data?	15
2.1 Inleiding	15
2.2 De 3,4,5,6,7 V's	15
2.3 Big Data technisch bezien	17
2.4 Big Data en een andere manier van denken	19
2.5 Conclusie	20
3 Het diagnostisch proces	21
3.1 Inleiding	21
3.2 Syndroomherkenning	21
3.3 Kennis over ziekten en aandoeningen	22
3.4 Diagnose door de computer	23
3.5 De computer genereert zijn eigen kennis	24
3.6 Diagnose door de computer met behulp van Big Data	24
3.7 Een behandelingsvoorstel door de computer	25
3.8 Betrouwbaarheid van wetenschappelijke publicaties	27
3.9 Kan Big Data de informatieleemte en betrouwbaarheid van beschikbare gegevens in de zorg verbeteren?	29
3.10 Conclusie	30
4 Het 'omics'-tijdperk	31
4.1 Inleiding	31
4.2 Next generation sequencing	32
4.3 Sequencing en Big Data	33
4.4 Farmacogenetica	35
4.5 Transcriptomics, proteomics, metabolomics	36
4.6 Aandachtspunten bij de nieuwe technologische ontwikkelingen	37
4.7 Vervagende grenzen tussen diagnostiek en onderzoek	38
4.8 Conclusie	38

5	Systeembioologie, Systems Medicine en P4 Medicine	39
5.1	Inleiding	39
5.2	Systems Medicine	39
5.3	P4 Medicine	41
5.4	Systems Medicine en P4 Medicine en overheidsbeleid	42
5.5	Conclusie	42
6	De private sector en Big Data in de zorg	45
6.1	Inleiding	45
6.2	Quantified Self	45
6.3	Health-platforms	46
6.4	Medisch onderzoek en private financiering	48
6.5	Conclusie	48
7	De omgeving: epidemiologie en Big Data	49
7.1	Inleiding	49
7.2	Relatie ziekte en omgevingfactoren	49
7.3	Bevolkingsonderzoeken en biobanken	50
7.4	Prevalentie en incidentie van (infectie)ziekten	53
7.5	Opsporen van bijwerkingen van geneesmiddelen	55
7.6	Voedselveiligheid	56
7.7	Het Internet of Things als bron voor (epidemiologische) gegevens	56
8	Big Data en kwaliteit en doelmatigheid van de zorg	59
8.1	Inleiding	59
8.2	Kwaliteit en veiligheid van zorg	59
8.3	Doelmatigheid van zorg	60
8.4	Conclusie	61
9	De gevaren en beperkingen van Big Data in de zorg	63
9.1	Inleiding	63
9.2	Big Data en privacy	63
9.3	Gegevensbeveiling	64
9.4	Informatie is macht	65
9.5	Beperkingen van Big Data	66
9.6	Conclusie	68
10	Beschouwing en toekomstperspectief van Big Data in de zorg	69
	Noten	73

VOORWOORD

WRR Working Paper 19 is geschreven als achtergrondstudie voor het project ‘Big Data, privacy en veiligheid’. In dit project verkent de WRR de implicaties van het gebruik van Big Data binnen het veiligheidsdomein.

De achtergrondstudie *Big Data in de zorg* is uitgevoerd door Leo Ottes, senior adviseur bij de Raad voor Volksgezondheid en Samenleving (RVS). Bij de RVS en diens voorgangers heeft Ottes meegewerkt aan de totstandkoming van een groot aantal adviezen en vele achtergrondstudies geschreven, zoals *Wetenschappelijke en technologische ontwikkelingen en nieuwe beroepen* (RVZ, 2011), *Sturen op gezondheidsdoelen* (RVZ, 2011), *Van Alzheimer tot Methusalem* (RVZ, 2012), *Theorie en praktijk van menselijk gedrag in een solidair zorgstelsel* (RVZ, 2013), Een internationale vergelijking voor het advies Patiënteninformatie (RVZ, 2014) en *Consumenten-eHealth: A game changer?!* (RVZ, 2015).

In *Big Data in de zorg* onderzoekt Ottes in hoeverre Big Data een bijdrage kan leveren aan de aanvulling van het grote tekort aan betrouwbare informatie in de zorg. Hij besteedt daarbij onder meer aandacht aan de bronnen voor Big Data-analyses, de gevoeligheid van zorggegevens en de rol van bedrijven als Apple, Google, Samsung en IBM, die de zorg als een aantrekkelijke groeimarkt beschouwen.

De serie ‘Working Papers’ omvat studies die in het kader van de werkzaamheden van de WRR tot stand zijn gekomen. De verantwoordelijkheid voor de inhoud en de ingenomen standpunten berust bij de auteurs.

Prof. dr. André Knottnerus
Voorzitter WRR

Dr. Frans Brom
Directeur WRR

SAMENVATTING

In deze verkenning wordt Big Data gezien als een ontwikkeling waarbij grote hoeveelheden gegevens, afkomstig uit verschillende gegevensbronnen, aan elkaar gerelateerd worden om te zoeken naar patronen zonder vooraf opgestelde hypothesen. De gegevens zijn vaak afkomstig uit verschillende domeinen en worden gebruikt voor andere doeleinden dan waarvoor ze oorspronkelijk verzameld zijn.

Big Data kan een belangrijke bijdrage leveren aan de aanvulling van het grote tekort aan betrouwbare informatie in de zorg. Alhoewel de kennis over ziekten en aandoeningen de afgelopen decennia sterk is toegenomen, is de precieze oorzaak, of het samenspel van oorzaken, van de meeste ziekten en aandoeningen niet bekend. Het overgrote merendeel kan niet écht genezen worden, hetgeen resulteert in chronische aandoeningen, die vaak alleen symptomatisch kunnen worden behandeld. Van de helft van de behandelingen is het effect onbekend, omdat dit niet onderzocht is. En van de helft die wel onderzocht is, is, zoals Ioannides heeft aangetoond, het resultaat veelal onbetrouwbaar.

Voor het opvullen van kennislacunes in de zorg kan aan verschillende 'Big Data-bronnen' gedacht worden, zoals gegevens over het genoom, die met behulp van zogenoemde *high-throughput*-technieken, zoals *next generation sequencing* relatief goedkoop verkregen kunnen worden of aan het *Internet of Things* (IoT). Bij dit laatste gaat het voor wat de zorg betreft voornamelijk om *lifestyle* gadgets, die onder andere de hoeveelheid lichaamsbeweging en de hartslag meten. Op het terrein van sensoren zijn er evenwel grote ontwikkelingen: ze worden steeds kleiner, goedkoper en kunnen steeds meer zaken meten, zoals ECG's, maar ook milieu-invloeden zoals hoeveelheden fijnstof en radioactiviteit. Een andere bron van gegevens vormen de medische dossiers, die inmiddels voor een groot deel gedigitaliseerd zijn. Daarnaast bevatten biobanken grote hoeveelheden waardevolle informatie in de vorm van bloed-, urine en weefselsamples. Integratie van al deze informatie uit verschillende bronnen zal de kennis over de oorzaak van ziekten en aandoeningen sterk doen toenemen. Op basis van deze toegenomen kennis kunnen effectieve(re) behandelingen ontwikkeld worden. De computer kan met al deze kennis gevoed worden – *artificial intelligence* – en het stellen van een diagnose en het kiezen van de beste behandeling ondersteunen. Big Data kan eveneens ingezet worden ter verbetering van de kwaliteit en doelmatigheid van de zorgverlening. Hiervan zijn reeds verschillende praktijkvoorbeelden.

Aan het gebruik van Big Data zijn evenwel ook beperkingen en gevaren verbonden. Onoordeelkundig gebruik kan tot onjuiste resultaten leiden. Zo kunnen met Big Data alleen correlaties tussen gegevens gevonden worden, maar geen causale

verbanden. Daarnaast zullen er veel correlaties gevonden worden die op toeval berusten of waarvan de relevantie niet bekend is. Voor een goede toepassing van Big Data is een gedegen kennis van de 'technische' beperkingen nodig.

Zorggegevens zijn zeer privacygevoelig. In het Big Data-tijdperk wordt het echter steeds lastiger om gegevens echt te anonimiseren, zodat ze voor wetenschappelijk onderzoek of beleid gebruikt kunnen worden. Door koppeling van gegevensbestanden kunnen geanonimiseerde gegevens vaak weer gedeanonimiseerd worden. Een belangrijke vraag is of privacy '*a thing of the past*' gaat worden. Burgers zeggen weliswaar zich zorgen te maken over hun privacy, maar in hun gedrag blijkt hier weinig van, gezien het gemak waarmee ze allerlei persoonlijke gegevens openbaar maken via sociale media. Veel 'gratis' diensten op het internet betaalt de gebruiker met zijn privacy. Privacy lijkt zich dan ook te verschuiven van iets waarop men recht heeft naar iets waarvoor men moet betalen. De vraag is of dit voor zorggegevens ook gaat gelden.

Een ander punt van zorg in relatie tot Big Data is het gevaar van centralisatie van menselijke kennis in casu van medische kennis. Grote bedrijven zoals Apple, Google, Samsung en IBM zien *Big Data* in de zorg als een aantrekkelijke groei-markt. Een aantal werkt aan systemen voor kunstmatige intelligentie op het terrein van medische diagnostiek en behandeling. Watson, de supercomputer van IBM, is het bekendste voorbeeld. Dergelijke systemen worden gevoed met wetenschappelijke kennis. De breed gedragen opvatting binnen de wetenschappelijke wereld is dat kennis vrijelijk voor iedereen beschikbaar moet zijn. Het is evenwel niet ondenkbaar dat er, net als bij de zoekmachines op het internet, één of enkele dominante partij(en) kom(t)(en), waardoor een centralisatie van kennis – een alomvattende *corpus of knowledge* – ontstaat die in private handen is. Daarnaast is er een ontwikkeling gaande waarbij ook de richting van kennisexploratie meer en meer privaat gestuurd wordt. Zo wordt op dit moment reeds veel klinisch onderzoek, met name *randomized controlled trials*, gefinancierd door de farmaceutische industrie. Zoals hiervoor aangegeven kan Big Data een belangrijke bijdrage leveren aan het opvullen van lacunes in kennis. Echter als deze gegevensstromen via private partijen lopen, dan zal er commerciële invloed zijn op de richting en wijze waarop de kennislacunes worden opgevuld.

1 INLEIDING

1.1 DE ZORG GENEREERT VEEL GEGEVENS

Binnen de (gezondheids)zorg worden reeds van oudsher grote hoeveelheden gegevens geproduceerd en verwerkt. In eerste instantie werden deze verzameld in het papieren patiënt/cliëntdossier. Deze bestond uit een bonte verzameling van handgeschreven verslagen, aantekeningen, prints van laboratoriumuitslagen, röntgenfoto's, enzovoort. Het grote probleem van het papieren dossier is, dat het weliswaar veel gegevens bevat, maar dat deze slecht toegankelijk zijn.

Het papieren dossier is inmiddels vrijwel overal vervangen door een elektronisch patiëntendossier. De informatie is daardoor op een meer gestructureerde wijze opgeslagen en veel beter toegankelijk. (Helaas geldt dat ook voor onbevoegden c.q. *hackers*, zoals verschillende inbraken in ziekenhuisinformatiesystemen in de Verenigde Staten laten zien.¹) Alle gegevens, laboratoriumuitslagen, röntgenfoto's, mri-, ct- en pet-scans enzovoort, worden inmiddels digitaal opgeslagen. Met name de beeldvormende technieken, zoals mri en ct, genereren veel bytes. In 2011 produceerde het Amerikaans gezondheidssysteem ca. 150 exabytes aan gegevens. Ter vergelijking: het totaal aantal woorden dat de mensheid tot op heden heeft uitgesproken wordt geschat op 5 exabytes.² De omvang van het elektronische medisch dossier, inclusief röntgenfoto's en dergelijke, van de Californische zorgverlener Kaiser Permanente ligt tussen de 26 en 44 petabytes.³ Voor de Nederlandse ziekenhuizen wordt het volume geschat op 2 petabytes (0,002 exabytes) per jaar.⁴ Ook in de zorg is een digitaliseringsslag gaande. Gegevens van cliënten worden meer en meer in elektronische cliëntendossiers vastgelegd. De totale omvang ervan is evenwel niet bekend.

De gegevens die worden vastgelegd in het elektronische patiënt/cliëntdossier worden nagenoeg alleen gebruikt voor de diagnostiek en behandeling van de betreffende patiënt/cliënt. Alhoewel de elektronische versie van het dossier beter toegankelijk is voor hulpverleners dan de papieren versie, staat het geautomatiseerd analyseren van de gegevens nog in de kinderschoenen. Het dossier bevat een variëteit aan gegevens. Veel gegevens zijn vastgelegd in vrije tekst, die lastig door de computer te interpreteren is. Hetzelfde geldt voor het beeldmateriaal.

1.2 KENNIS VAN OORZAKEN VAN ZIEKTEN EN AANDOENINGEN IS BEPERKT

Een groot probleem in de geneeskunde is dat ondanks de grote hoeveelheid kennis die er reeds beschikbaar is, de 'echte' oorzaken van ziekten en aandoeningen van met name de frequent voorkomende aandoeningen, zoals hart- en vaatziekten,

diabetes of kanker, niet bekend zijn. Er zijn wel veel risicofactoren bekend, die de kans op deze ziekten vergroten, maar in verreweg de meeste gevallen gaat het om vrij onspecifieke zaken zoals voeding of te weinig bewegen. Maar het exacte mechanisme waarlangs bijvoorbeeld bewegen het ontstaan van hart- en vaatziekten beïnvloedt, is niet bekend. Hierdoor is de behandeling van de meeste aandoeningen vooralsnog symptomatisch en veelal een proces van *trial-and-error*. Zo bestaat de gangbare behandeling bij een vernauwing van de kransslagader van het hart uit het ‘platduwen’ van de atherosclerotische plaque met een ballonnetje en het openhouden van het vat met een stukje metaalgaas. Een andere optie is het aanbrengen van een omleiding rond de vernauwing. Echter voor de atherosclerose zelf bestaat geen afdoende behandeling.

Voor vrijwel alle aandoeningen geldt hetzelfde verhaal, of het nu diabetes, reumatoïde artritis, MS of dementie betreft: de symptomen en de progressie van de aandoening zijn vaak te behandelen, maar echt genezen lukt niet. In dat opzicht is de term *geneeskunde* dan ook ongelukkig gekozen.

1.3 KWALITEIT EN DOELMATIGHEID VAN DE ZORG

De stijgende kosten van de gezondheidszorg staan reeds jaren op de politieke agenda. Een veelgestelde vraag is of de kosten in verhouding staan tot de baten. Wat levert de zorg op in termen van gezondheidswinst? Harde gegevens ontbreken vooralsnog voor een groot deel van de geleverde zorg. Er wordt wel getracht in deze leemte te voorzien, onder andere in de vorm van *evidence-based medicine* (EBM). Deze ontwikkeling wordt echter bekritiseerd. Zorgverleners moeten hun handelen steeds meer verantwoorden. Dit leidt tot een zware administratieve belasting die ten koste gaat van de patiëntenzorg. Daarnaast is er fundamentele kritiek op de wijze waarop EBM op dit moment wordt ingevuld. Met name de grote nadruk op ‘*hard evidence*’ in de vorm van *randomized controlled trials* (RCT’s) ligt onder vuur.

1.4 BIG DATA TO THE RESCUE?

Door ontwikkelingen op verschillende technologische terreinen en met name door het combineren van verschillende technologieën neemt de wetenschappelijke kennis sterk toe. Belangrijke technologische ontwikkelingen zijn de zogenoemde *high-throughput*-technieken, zoals *next generation sequencing*, waarbij in korte tijd en tegen relatief lage kosten grote hoeveelheden analyses verricht kunnen worden. Dit heeft geleid tot allerlei nieuwe onderzoeksvelden, zoals *genomics* (het in kaart brengen van alle genen) en *proteomics* (het in kaart brengen van alle eiwitten in een cel). Aangezien de namen van deze onderzoeksvelden allemaal eindigen op ‘*omics*’, spreekt men wel van het ‘*omics*’-tijdperk⁵, waarbij enorme hoeveelheden data gegenereerd worden. De interactie tussen de verschillende ‘*omics*’-

niveaus is zeer complex, maar door de grote hoeveelheden gegevens die de nieuwe onderzoeksmethoden opleveren aan elkaar te relateren en de onderlinge verbanden in de vorm van netwerken in kaart te brengen, tracht men deze interactie te ontrafelen. De tak van de biologie die zich hiermee bezighoudt wordt de netwerk- of systeembioogie genoemd. Bij de toepassing van deze aanpak binnen de geneeskunde spreekt men van *Systems Medicine*.

Om de onderliggende mechanismen van gezondheid en ziekte te kunnen doorgronden, moeten de gegevens over genetische aanleg en dergelijke naast omgevingsfactoren worden gelegd. Onderzoek van de relatie tussen gezondheid en ziekte enerzijds en omgevingsfactoren anderzijds is het terrein van de epidemiologie. Het gaat hierbij om een heel palet aan factoren, van dichtbij of zelfs in het lichaam, zoals de verschillende bacteriepopulaties op verschillende plaatsen op de huid en in de darmen (microbioom), en chemische stoffen in het milieu tot psychische en sociale factoren.

1.5 LEESWIJZER

In deze notitie wordt het hiervoor beschrevene nader uitgewerkt. De term Big Data wordt vaak als synoniem gezien voor ‘heel veel gegevens’. Big Data is evenwel veel meer. Het is ook een manier van denken en een terrein waarop snelle ontwikkelingen gaande zijn. In hoofdstuk 2 wordt hierop nader ingegaan. De problematiek rond de huidige diagnostiek en behandeling van ziekten en aandoeningen en de (on)mogelijkheden die Big Data biedt voor de oplossing daarvan wordt in hoofdstuk 3 nader uitgewerkt. In hoofdstuk 4 wordt het ‘*omics*’-tijdperk, waarbij Big Data een cruciale rol speelt, nader beschreven. De systeembioogie tracht de *Big Data* die op de verschillende terreinen wordt gegenereerd met elkaar in verband te brengen. Dit komt in hoofdstuk 5 aan de orde. Hoofdstuk 6 gaat nader in op het bedrijfsleven dat Big Data in de zorg als een interessante groeiemarkt ziet. De epidemiologie in relatie tot Big Data wordt in hoofdstuk 7 besproken. In hoofdstuk 8 worden de mogelijkheden gezien waarin Big Data kan bijdragen aan een betere en doelmatiger zorgverlening. Hoofdstuk 9 gaat nader in op de schaduwzijde van het gebruik van Big Data in de zorg, met name de privacyproblematiek, alsmede de beperkingen van Big Data. Tot slot wordt in hoofdstuk 10 getracht een toekomstperspectief te schetsen: hoe zal Big Data in de zorg er over 20 jaar uitzien?

2 WAT IS BIG DATA?

2.1 INLEIDING

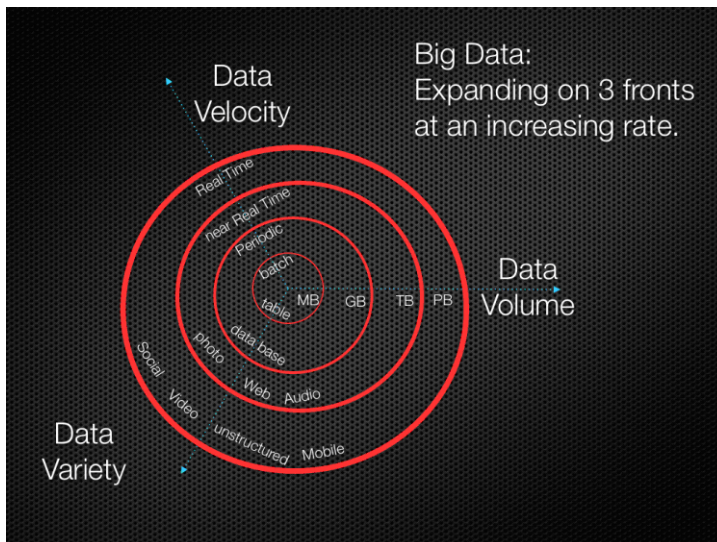
Big Data staat op dit moment in het middelpunt van de belangstelling. Het is een *hype* binnen de ICT, waar velen op in willen haken. Deze populariteit leidt ertoe dat de term te pas en onpas wordt gebruikt. Zoals uit het navolgende blijkt kan Big Data vanuit verschillende invalshoeken gezien en gedefinieerd worden.

2.2 DE 3,4,5,6,7 V'S

Een eerste vereiste om gegevens Big Data te kunnen noemen is dat het moet gaan om een grote hoeveelheid, maar dat is niet voldoende. Laney beschreef in 2001 Big Data in termen van 'de 3 V's': *Volume*, *Velocity* en *Variety*.⁶ *Volume* heeft betrekking op de grootte van de gegevensverzameling, *velocity* op de snelheid waarmee de gegevens binnenkomen en verwerkt moeten worden en *variety* op de variëteit van de gegevens, bijvoorbeeld ongestructureerde tekst, geluid, beelden en door sensoren gemeten fysische grootheden.

Soubra illustreert dit in een diagram, zie onderstaand figuur7.

Figuur 2.1 De drie V's van Big Data



Bron: Datasciencecentral⁸

Op de drie assen is de ontwikkeling van volume, snelheid en variëteit in de loop der tijd aangegeven. In het begin van het computertijdperk werden gegevens in tabelvorm *batch*-gewijze verwerkt. De omvang van de gegevensverzamelingen was in de orde van grootte van megabytes ($1 \text{ MB} = 2^{10} \text{ Kilobytes} = 2^{20} \text{ bytes} = 1.048.576 \text{ bytes}$, waarbij een byte bijvoorbeeld een letter of een cijfer kan representeren). Vervolgens kwamen er databases in de orde van grootte van gigabytes ($1 \text{ Gb} = 1024 \text{ MB} = 2^{30} \text{ bytes}$). Met de opkomst van het internet nam de variëteit van de gegevens toe, met digitale audio en foto's. Het volume steeg tot terabytes ($1 \text{ TB} = 2^{40} \text{ bytes}$). De ontwikkeling van snel internet met *streaming video*, *social media* en mobiele communicatie hebben dit alles in een verdere stroomversnelling gebracht, met gegevensvolumes in de orde van grootte van petabytes ($1 \text{ PB} = 10^{50} \text{ bytes}$). (Om het lijstje van gegevensvolumes te completeren: men spreekt van exabyte (EB) in geval van 2^{60} bytes, zettabyte (ZB) bij 2^{70} bytes en yottabyte (YB) bij 2^{80} bytes.)

Er is geen exacte grens aan te geven voor de grootte van de verschillende V's waarboven sprake is van Big Data. Zo kunnen bijvoorbeeld de sensoren van de Large Hadron Collider, waarmee fundamenteel fysisch onderzoek wordt verricht, per dag 500 exabytes genereren, meer dan 200 keer zoveel als alle andere gegevensbronnen in de wereld bij elkaar. Slechts een klein deel van deze gegevens, 0,001% kan daadwerkelijk verwerkt worden, hetgeen in 2012 25 petabytes aan gegevens opleverde.⁹ Ondanks deze duizelingwekkende hoeveelheid gegevens (en snelheid van registratie), is het volgens de beschrijving van Laney geen Big Data, omdat variëteit in de gegevens ontbreekt.

Snelheid van gegevensverwerking is eveneens belangrijk. Zo zijn reeds enkele decennia modellen beschikbaar om het weer te voorspellen. Echter tot in de jaren negentig van de vorige eeuw duurde het tot wel vijftien dagen om een weersvoorspelling voor de komende zeven dagen door te rekenen. Heden ten dage kan dit in enkele uren, waardoor het een praktisch bruikbaar instrument is geworden.

Het is een soort sport geworden om aanvullende V's te bedenken bij Big Data. Zo heeft IBM een vierde V toegevoegd, namelijk *veracity* – waarheidsgetrouwheid.¹⁰ Gegevens van de ene gegevensbron zijn betrouwbaarder dan die afkomstig van een andere. Bij de verwerking moet hiermee rekening worden gehouden. Het Institute for Health Technology Transformation in de VS geeft aan dat met name voor de zorg waarheidsgetrouwheid belangrijk is.¹¹ Een belangrijke vraag is of gegevens met een lage waarheidsgetrouwheid, zogenoemde '*dirty data*', getransformeerd kunnen worden in '*clean data*'. Vaak wordt als kracht van Big Data genoemd dat daarin juist allerlei fouten, onzuiverheden enzovoort zich uitmiddelen, met als resultaat betrouwbare gegevens. Bij systematische afwijkingen lukt dit echter niet. In paragraaf 9.5 wordt nader ingegaan op de beperkingen van Big Data.

Andere voorgestelde V's zijn *variability*, waarmee de inconsistentie tussen gegevens wordt bedoeld, *value*, de waarde van de gegevens, *visualisation*, het begrijpelijk presenteren van de gegevens, en tot slot *volatility*, de tijdelijkheid c.q. korte houdbaarheid van de gegevens.^{12,13}

2.3 BIG DATA TECHNISCH BEZIEN

Het begrip Big Data kan ook worden gezien vanuit technologisch oogpunt. De toegenomen snelheid van gegevensverzameling is te danken aan de ontwikkeling van zeer snelle en goedkope internetverbindingen, met name glasvezelverbindingen. Daarnaast is de opslagcapaciteit voor gegevens sterk toegenomen tegen een steeds lagere prijs. Zo kostte de opslag van 20 GB aan gegevens in 1980 1 miljoen dollar. Het daarvoor benodigde apparaat was een vijftal archiefkasten groot en woog circa 1,5 ton. Tegenwoordig kost een USB-stick of geheugenkaartjes van 32 GB minder dan 20 euro en weegt slechts enkele tientallen grammen.¹⁴ Hetzelfde geldt voor de gegevensverwerking. Gegevens kunnen steeds sneller verwerkt worden door snellere en goedkopere processoren. Door meerdere processoren gelijktijdig gegevens te laten verwerken -*parallel processing*- kan de snelheid nog verder opgevoerd worden.¹⁵

Databases

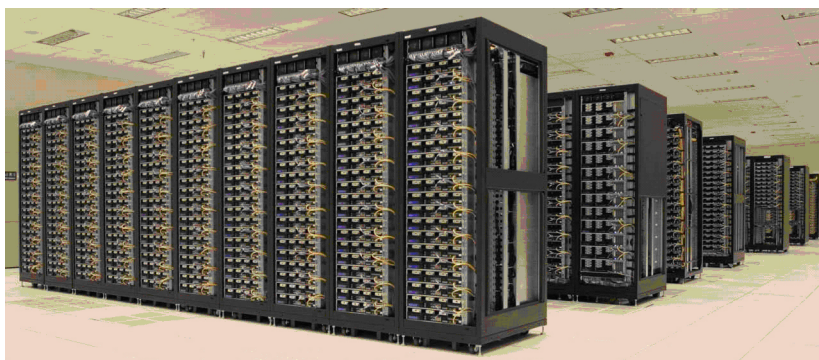
In het verleden waren relationele *database management* systemen (RDBMS), ook wel SQL-databases genoemd, naar de gebruikte vraagtaal *Structured Query Language*, de standaard voor de verwerking van wat toen gezien werden als grote hoeveelheden gegevens. Het probleem van deze databases is de schaalbaarheid. SQL-databases kunnen op een gegeven moment het almaar toenemende volume van gegevens niet meer binnen redelijke tijd verwerken. Vanuit dit probleem wordt Big Data ook wel gedefinieerd als: "Gegevensverzamelingen die zo groot en complex zijn dat traditionele gegevensverwerkende instrumenten, zoals relationele databases, niet in staat zijn om deze binnen acceptabele tijd en kosten te verwerken."¹⁶

Een van de eersten die daar tegenaan liep, was de in de jaren negentig van de vorige eeuw zeer populaire internetzoekmachine Yahoo!. Een kernactiviteit van een zoekmachine is het indexeren van het World Wide Web. De groei hiervan maakte dat conventionele strategieën op een gegeven moment niet meer werkten. Yahoo! had een computerplatform nodig dat altijd dezelfde hoeveelheid tijd nodig had om het Web te indexeren, ongeacht de grootte. Aangezien zij de oplossing op de commerciële markt niet kon vinden, ontwikkelde zij daarvoor zelf een computerprogramma, *Hadoop* genaamd, voor haar netwerk van tienduizenden dataservers en stelde het programma als open-source beschikbaar.¹⁷ Dit betekent dat de broncode voor iedereen vrijelijk beschikbaar is. Inmiddels is het programma geadopteerd door de internetcommunity en de Apache Software Foundation heeft een belang-

rijke rol gespeeld bij de verdere ontwikkeling. Het ‘Hadoop-platform’ biedt inmiddels een heel palet aan gratis, *open-source*, softwaregereedschappen voor het verwerken van Big Data.¹⁸

De Hadoop-systemen voor de opslag en verwerking van Big Data hebben een aantal specifieke kenmerken. Zo worden gegevens niet op één systeem opgeslagen, maar verdeeld over vele dataservers. Het geheel is ‘*scalable*’, hetgeen betekent dat het gemakkelijk uitgebreid kan worden tot duizenden servers verdeeld over één of meerdere datacenters, zie onderstaand figuur.

Figuur 2.2 **Voorbeeld van een datacenter**



Bron: Yahoo! Hadoop cluster [OSCON '07]

De servers bestaan uit (relatief) goedkope standaard onderdelen, zoals processoren en harde schijven (tegenwoordig ook steeds meer ‘niet-mechanische harde schijven’, zogenoemde *solid state drives* afgekort SSD). Apparaten kunnen defect raken. Bij gebruik van duizenden servers is er altijd wel iets dat hapert. Dit is vaak de mechanische harde schijf waarop de gegevens opgeslagen worden. Dezelfde gegevens worden daarom op meer dan één plek – redundant – opgeslagen, zodat het systeem zichzelf kan herstellen. De gegevens worden door vele servers tegelijkertijd verwerkt op de locatie waar de gegevens worden of zijn opgeslagen: *local parallel processing*. In de praktijk spreekt men dan ook van ‘Hadoop-clusters’. Is er meer opslag en verwerkingscapaciteit nodig, dan koppelt men simpelweg extra servers met reken- en opslagcapaciteit aan. Hadoop-clusters kunnen gezien worden als goedkope, schaalbare supercomputers.¹⁹ Door de verkoop van computertijd op Hadoop-clusters zijn supercomputerfaciliteiten inmiddels binnen het bereik van kleinere organisaties en particulieren gekomen.²⁰

2.4 BIG DATA EN EEN ANDERE MANIER VAN DENKEN

Zoals in het voorgaande is aangegeven, kan Big Data vanuit verschillende invalshoeken gezien en gedefinieerd worden. Het mag geen verwondering wekken dat de Nationale DenkTank in haar eindrapport Big Data in zicht uit 2014 concludeert, dat er geen overeenstemming over de definitie is en dat er dus geen harde grens is die stelt wat wel of niet onder Big Data valt. Zij kiest daarom voor een pragmatische definitie, waarbij *Big Data* gezien wordt als een ontwikkeling: “Big Data is de enorme toename van mogelijkheden om data te genereren, combineren en analyseren, die leidt tot nieuwe inzichten en een nieuwe manier van redeneren.”²¹

Ook de WRR onderkent dat er geen overeenstemming over een eenduidige definitie van Big Data bestaat en ziet het meer als een samenspel van ontwikkelingen. Deze notitie sluit zich bij deze visie aan. Daarbij kan een aantal kenmerken genoemd worden waarbij een onderscheid gemaakt kan worden in de aard van de data, de analysetechnieken en het gebruik:

Aard van de data

Er is sprake van grote hoeveelheden gegevens die vaak on- of semi-gestructureerd zijn. Daarnaast zijn ze gevarieerd en veelal afkomstig uit verschillende databronnen.

Analysetechnieken

De analysetechnieken zijn *data driven*. Dit betekent dat naar patronen gezocht wordt zonder dat vooraf hypothesen zijn opgesteld zoals bij de klassieke aanpak. De resultaten zijn snel beschikbaar en geven daardoor inzicht in het nu: (*near*) *real-time* c.q. *nowcasting*. Ze kunnen voorspellende waarde hebben: *predictive analysis/forecasting*.

Gebruik

De verschillende gegevensbronnen bevinden zich veelal in verschillende domeinen. Dit leidt tot ontschotting van domeinen: gegevens uit het ene domein worden gebruikt voor beslissingen in het andere domein. De resultaten op geaggregeerd niveau kunnen toegepast worden op beslissingen op groeps- of individueel niveau.

In de gezondheidszorg zijn bij al deze kenmerken grote ontwikkelingen gaande. Zo leveren bijvoorbeeld nieuwe *next generation sequencing*-technieken enorme hoeveelheden genetische gegevens die men tracht te relateren aan allerlei andere gegevens, van voeding en gevoeligheid voor luchtvervuiling tot psychische stoornissen. Het gaat daarbij vaak om gegevens die voor een ander doel en binnen een ander domein zijn vastgelegd. Door het op de markt komen van allerlei kleine goedkope sensoren kunnen steeds meer fysiologische parameters, zoals ECG,

bloeddruk, geautomatiseerd en continu geregistreerd worden. De gegevens kunnen heel eenvoudig draadloos, bijvoorbeeld via een *smartphone*, in grote gegevensbestanden vastgelegd worden en gerelateerd worden aan de andere gegevens om hierin naar patronen te zoeken.

Globaal kunnen vijf categorieën Big Data in de zorg onderscheiden worden:²²

1. *'Human generated data'*: Door zorgverleners in het elektronisch medisch dossier vastgelegde gegevens, aantekeningen, e-mails, ontslagbrieven, enzovoort. Veel van deze gegevens zijn ongestructureerd of semi-gestructureerd.
2. Transactiegegevens: declaraties, enzovoort.
3. Biometrische gegevens. Hieronder vallen naast vingerafdrukken en netvlies-scans ook bijvoorbeeld röntgenfoto's, CT- en MRI-scans en fysiologische meetgegevens, zoals bloeddruk, hartslag, zuurstofverzadiging van het bloed, enzovoort.
4. *Machine-to-machine-data*: bijvoorbeeld door sensoren gegenereerde gegevens die door bewakingsapparatuur worden verwerkt.
5. Websites en *social media*: deze omvatten bijvoorbeeld klik- en interactiegegevens ('likes' en dergelijke) van sociale media zoals Facebook, Twitter, LinkedIn, blogs, *smartphone apps* en dergelijke.

2.5 CONCLUSIE

In dit hoofdstuk is nader ingegaan op de vraag wat Big Data nu eigenlijk is. Het blijkt dat hier geen eenduidig antwoord op te geven is. Big Data is meer een samenspel van ontwikkelingen. Technologische ontwikkelingen maken het mogelijk om grote hoeveelheden (gedigitaliseerde) gegevens razendsnel via de datasnelweg c.q. het internet tussen computers uit te wisselen, op te slaan, te ontsluiten en te bewerken. Daarbij is er een grote variëteit aan gegevens, namelijk alles wat gedigitaliseerd kan worden: getallen, tekst, geluid, video, enzovoort. Deze vormen de drie V's van Big Data: *volume*, *velocity* en *variety*.

De technologie biedt mogelijkheden voor nieuwe toepassingen c.q. nieuwe analysetechnieken van gegevens. Terwijl vroeger gegevens werden verzameld op basis van vooraf gestelde hypothesen kan nu in grote hoeveelheden gegevens naar patronen gezocht worden. Een belangrijk kenmerk daarbij is dat gegevens uit verschillende domeinen aan elkaar gekoppeld kunnen worden. Daarbij worden gegevens vaak voor een geheel ander doel gebruikt dan waarvoor ze oorspronkelijk verzameld werden.

3 HET DIAGNOSTISCH PROCES

3.1 INLEIDING

Om een ziekte of aandoening adequaat te kunnen behandelen moet een arts eerst de juiste diagnose stellen: aan welke ziekte of aandoening lijdt de patiënt? De diagnose vormt het handvat voor de keuze van de behandeling. De kennis over wat de beste behandeling is voor een bepaalde ziekte of aandoening, is veelal via *trial-and-error* tot stand gekomen door de uitkomsten van verschillende behandelingen met elkaar te vergelijken. In dit hoofdstuk wordt gekeken of Big Data hierin verandering kan brengen.

3.2 SYNDROOMHERKENNING

In veel gevallen heeft de patiënt vage klachten, zoals moeheid of buikpijn, of specifieke symptomen zoals koorts. Elk van deze symptomen kunnen bij verschillende aandoeningen voorkomen, maar de kern van de diagnostiek is de aanname dat een specifieke aandoening een specifieke combinatie van symptomen heeft. Het is de kunst om deze combinatie van symptomen – een syndroom – te herkennen. Het gaat hier om patroonherkenning, iets waar het menselijk brein uitermate goed in is. De arts vergelijkt hierbij zijn of haar kennis van bekende ziektepatronen met het patroon dat de patiënt presenteert. Het resultaat kan zijn dat de arts het patroon direct herkent (of meent te herkennen), deels herkent of niet herkent. Dit proces is grotendeels intuïtief en snel, vaak binnen dertig seconden. De mens is (vooralsnog) superieur in het herkennen van patronen ten opzichte van de computer.

Inductief en deductief redeneren

Op basis van het (deels) herkende patroon stelt de arts een initiële werkhypothese op. Dit is het inductieve gedeelte van het redeneringsproces: vanuit de symptomen wordt een mogelijke diagnose overwogen. Vervolgens wordt door middel van vervolgonderzoek getracht de werkhypothese te bevestigen of te verwerpen. Dit is het deductieve gedeelte: veronderstel dat de diagnose correct is, welke symptomen ‘behoren dan bij de aandoening voor te komen’? Veronderstel dat de arts denkt aan een ontsteking. Bekend is dat bij een ontsteking ‘de bezinking’ in het bloed verhoogd is. De arts zal vervolgens laboratoriumonderzoek aanvragen ter ondersteuning of verwerping van de hypothese ‘ontsteking’.

Een probleem bij het inductief redeneren is dat het in de praktijk lastig blijkt een hypothese te verwerpen op basis van nieuwe tegenstrijdige informatie. De mens is geneigd eerder te zoeken naar informatie die een eerste indruk ondersteunt, dan te zoeken naar informatie die deze ontkracht. Met spreekt van *anchoring* respectievelijk

lijk *confirmation bias*.^{23,24,25} Het verdisconteren van nieuw informatie vraagt gedegen kennis en vaardigheid voor wat betreft kansberekening. Hoe de kans op een bepaalde aandoening verandert in het licht van een reeks laboratoriumuitslagen blijkt lastig in te schatten. Ook is men al snel tevreden met het resultaat en zoekt men niet verder (*premature closure*)²⁶. Dit kan leiden tot tunnelvisie.

Deze fenomenen zijn ook artsen niet vreemd.^{27,28} Uit onderzoek is gebleken dat artsen vaak aan hun initiële hypothese vasthouden, ook als er contradictoire informatie beschikbaar is gekomen.^{29,30}

Bij het deductief redeneren is het probleem dat de kennis van de individuele arts – het aantal patronen dat de arts ‘in zijn hoofd heeft’ – beperkt is. Het is niet voor niets dat er medische specialisaties zijn. Eén individu kan onmogelijk alle aandoeningen herkennen. Daar komt bij dat aandoeningen die men infrequent ziet naar de achtergrond geraken, waardoor het moeilijker wordt deze te herkennen. Dit leidt tot een *bias* voor frequent en recent geziene aandoeningen. Een bekend probleem is het herkennen van zeldzame aangeboren aandoeningen. Het duurt bij een kwart van de patiënten vijf tot dertig jaar voordat de juiste diagnose wordt gesteld.³¹

3.3 KENNIS OVER ZIEKTEN EN AANDOENINGEN

Het diagnostisch proces zoals in het voorgaande geschetst, is gebaseerd op ziektebeelden en gaat uit van de veronderstelling dat er ‘op zichzelf staande’ ziekten en aandoeningen bestaan, zoals bijvoorbeeld reumatoïde artritis. Echter het onderkennen van afgrensbare ziekten en aandoeningen met specifieke symptomen en kenmerken is, net zoals het diagnostisch proces, mensenwerk en kent zijn beperkingen. Zolang de precieze oorzaak van een aandoening niet bekend is, bestaat de mogelijkheid dat wat nu als verschillende aandoeningen worden gezien in werkelijkheid een en dezelfde aandoening is, die zich bijvoorbeeld door genetische verschillen, anders uit bij verschillende patiënten.

Een andere illustratie is de discussie rond de recent herziene DSM5 (Diagnostic and Statistical Manual of Mental Disorders 5th Edition)³², waarin wordt aangegeven wat onder een bepaalde psychiatrische aandoening moet worden verstaan. Over de wetenschappelijke basis voor de verschillende te onderscheiden ziektebeelden bestaat grote twijfel.³³

3.4 DIAGNOSE DOOR DE COMPUTER

Reeds vanaf de jaren 60 van de vorige eeuw poogt men kennis van deskundigen ‘af te tappen’ en in een computer te stoppen. Dit is het terrein van de kunstmatige intelligentie. Eén van de eerste programma’s op het terrein van de geneeskunde was MYCIN, zie kadertekst.

Kunstmatige intelligentie: MYCIN

MYCIN was een van eerste de toepassingen van kunstmatige intelligentie in de geneeskunde. Het is in de jaren ‘60 van de vorige eeuw ontwikkeld in het Stanford Medical Center en is gebaseerd op de concepten van inductief redeneren.³⁴ Het programma was gericht op het diagnosticeren en behandelen van bacteriële infecties met antibiotica. MYCIN kon in 69% van de gevallen de juiste therapie selecteren. Nu lijkt dit percentage niet zo hoog, doch het programma deed het beter dan de gemiddelde arts.

Bronnen: Wikipedia en Buchyanan en Shortcliffe³⁵

Ondanks dat MYCIN het beter deed dan de gemiddelde arts, is het systeem nooit in de dagelijkse praktijk gebruikt. De belangrijkste reden was dat de technologie er nog niet klaar voor was. In de jaren zeventig bestond de *personal computer* nog niet. De gebruiker moest de antwoorden op de vragen die MYCIN stelde intypen, waarna het programma op een *time-shared* systeem ruim een half uur nodig had om tot een antwoord te komen.

Tegenwoordig zijn deze technologische problemen grotendeels opgelost. Toch zijn expertsystemen zoals MYCIN nooit echt doorgebroken in de dagelijkse praktijk. Het grote probleem bij de ontwikkeling van expertsystemen bleek het aftappen van de kennis van de expert, de zogenoemde *knowledge acquisition bottleneck*. De kennis moet opgeslagen worden in een *knowledgebase* in de vorm van *decision-rules*. De *knowledgebase* van MYCIN bestond uit circa 600 *rules*.³⁶

Zoals in het voorgaande is aangegeven, is de kennis in het brein voornamelijk opgeslagen in patronen en het is voor de expert moeilijk om precies aan te geven hoe hij of zij tot een conclusie komt.

Een belangrijk probleem bij het stellen van een diagnose door de computer is de juridische aansprakelijkheid. De arts is immers eindverantwoordelijk. Een ander punt is of een computerdiagnose ook daadwerkelijk geaccepteerd wordt. Als uit wetenschappelijk onderzoek zou blijken dat de computer beter diagnoses kan stellen, willen we er dan op vertrouwen?

3.5 DE COMPUTER GENEREERT ZIJN EIGEN KENNIS

Een mogelijk oplossing voor het probleem dat experts hun kennis moeilijk expliciet kunnen maken is om geen gebruik te maken van hun kennis, maar de computer op basis van vooraf geformuleerde algoritmen uit grote databases met medische gegevens correlaties te laten vinden tussen aandoeningen, klachten, symptomen, labwaarden en andere onderzoeksuitslagen. Het gaat hierbij om grote hoeveelheden verschillende soorten gegevens, een uitdaging bij uitstek voor Big Data.

Op basis van de gegevensanalyses kunnen de correlaties tussen aandoeningen enerzijds en klachten, symptomen en onderzoeksresultaten anderzijds in de vorm van kansen bepaald worden. Een ziekte of aandoening c.q. syndroom bestaat dan uit een aantal bij elkaar horende klachten, symptomen en onderzoeksuitslagen tezamen met voor elke klacht, symptoom of onderzoeksuitslag de kansen dat deze optreden bij de betreffende aandoening.

Privacy van de patiënt

Een probleem bij deze aanpak is de privacy van de patiënt. De computer moet immers grootschalige toegang hebben tot patiëntgegevens. In hoofdstuk 9 wordt de problematiek rond privacy nader besproken.

3.6 DIAGNOSE DOOR DE COMPUTER MET BEHULP VAN BIG DATA

Als voor elke aandoening de kans op het optreden van een bepaalde klacht, symptoom of laboratoriumuitslag bekend is, kan de computer met behulp van Bayesiaanse statistiek, zie kadertekst, de kansen op verschillende aandoeningen berekenen gegeven de klachten, symptomen en onderzoeksresultaten van de patiënt en zo een differentiaaldiagnose opstellen, waarbij de meest waarschijnlijke diagnose bovenaan staat.

Bayesiaanse statistiek

Veronderstel iemand is bang dat hij een zeldzame ziekte heeft en laat zich hiervoor testen. Deze zeldzame ziekte komt in de bevolking bij 1 op de 10.000 mensen voor en de betreffende test geeft in 99% van de gevallen de juiste uitslag. (Dit betekent dat als 100 personen de ziekte hebben, de test bij 99 positief is. Dus 1 persoon die de ziekte wel heeft wordt gemist. En omgekeerd: bij 100 personen die de ziekte niet hebben, valt de test bij 99 personen negatief uit. Dus 1 gezonde persoon wordt ten onrechte als ziek aangemerkt.)

De testuitslag blijkt positief. Wat is nu de kans dat de persoon daadwerkelijk de zeldzame ziekte heeft? Intuïtief denken de meeste mensen dat deze kans hoog is. De test geeft immers in 99% van de gevallen de juiste uitslag. In werkelijkheid is de kans echter minder dan 1%, namelijk 0,0098.

Deze kans kan berekend worden met behulp van het theorema van Bayes, dat stamt uit 1763, en luidt:

$$P(A|B) = P(A) \cdot P(B|A) / P(B)$$

Hierbij zijn $P(A)$ en $P(B)$ de kansen op gebeurtenis A resp. B. In het voorbeeld is gebeurtenis A dat iemand de ziekte heeft en gebeurtenis B een positieve testuitslag. De kans dat iemand de ziekte heeft zonder dat de uitslag van de test bekend is, bedraagt 1 op 10.000 ($=0,0001$); de kans op de ziekte in de hele populatie. De kans dat iemand een positieve testuitslag krijgt is gelijk aan de som van de kans op een terecht positieve en onterecht positieve uitslag. De kans op een terecht positieve uitslag is gelijk aan de kans dat de testuitslag positief is gegeven dat de persoon de ziekte heeft maal de kans dat de persoon de ziekte heeft, dus $0,99 * 0,0001 = 0,000099$. De kans op een onterecht positieve uitslag is de kans dat de test toch positief is terwijl de persoon de ziekte niet heeft maal de kans dat de persoon de ziekte heeft, dus $0,01 * 0,9999 = 0,009999$. $P(B) = 0,000099 + 0,009999 = 0,010098$ $P(A|B)$ is dan de kans dat iemand de ziekte daadwerkelijk heeft gegeven een positieve testuitslag. Dat is hetgeen we willen weten:
 $P(A|B) = 0,0001 * 0,99 / 0,010098 = 0,0098039$ (0,98%)

Bron: Harver Mudd College Math Department³⁷

Vervolgens kan de computer aangeven welk vervolgonderzoek het best geschikt is. Met de resultaten van het vervolgonderzoek kan de computer alles opnieuw doorrekenen en een aantal aandoeningen uitsluiten of juist bevestigen en zo tot een verdere verkorting van de lijst van mogelijke aandoeningen komen. Dit proces kan zich herhalen totdat de meest waarschijnlijke diagnose gesteld kan worden.

Bij de analyse kan de computer tevens rekening houden met de kosten die verschillende onderzoeken met zich meebrengen en zo het meest kosteneffectieve pad berekenen om tot een diagnose te komen.

Het voordeel van de computer is dat deze geen last heeft van eerder beschreven menselijke problemen zoals *anchoring* en *confirmation bias* en *premature closure*. Een computer staat indifferent tegenover verschillende hypothesen. Het berekent enkel kansen.

3.7 EEN BEHANDELINGSVOORSTEL DOOR DE COMPUTER

Als de diagnose gesteld is, kan de computer vervolgens in alle gepubliceerde digitale informatie zoeken naar de meest geschikte behandeling gegeven de situatie van de patiënt. Om dit mogelijk te maken zijn in elke geval twee zaken nodig: een krachtige computer die de grote hoeveelheid gegevens snel kan verwerken en de beschikbaarheid van voldoende relevante gegevens over de effectiviteit van behandelingen. Op dit moment wordt daar reeds druk aan gewerkt. De supercomputer Watson van IBM kan zich op dit moment op het terrein waarop het spel Jeopardy zich afspeelt reeds meten met de beste menselijke breinen, zie kadertekst. De kracht waarmee computers gegevens kunnen verwerken neemt nog steeds toe. Google's *director of Engineering*, Ray Kurzweil, verwacht dat de computer het menselijk brein in 2029 voorbij zal streven.³⁸ Overigens heeft Google zich inmiddels ook op het terrein van de kunstmatige intelligentie begeven door de aankoop van het daarin gespecialiseerde bedrijf DeepMind begin 2014. Later in dat jaar kocht

Google nog twee *startups* op dit terrein, DarkBlue Labs en Vision Factory.³⁹ Google is inmiddels een belangrijke speler op dit terrein. In oktober 2015 versloeg met het programma Alpha Go van Google voor de eerste maal in de geschiedenis overtuigend een professionele speler van het spel Go, namelijk de Europees kampioen.⁴⁰ Go is net als schaken een bordspel, maar met veel meer mogelijkheden, meer dan het aantal atomen in het heelal. Net als Watson is het een zelflerend systeem. Het maakt gebruik van een zogeheten *deep neural network*. De overwinning wordt gezien als een doorbraak op het terrein van de kunstmatige intelligentie. Het is naar verwachting slechts een kwestie van tijd dat de wereldkampioen Go zal worden verslagen en de mens, net als nu reeds bij het schaakspel, geen kans meer maakt tegen de computer.

Supercomputer Watson

De supercomputer Watson van IBM, genoemd naar de oprichter van het bedrijf, is vooral bekend geworden door het winnen van het spel Jeopardy op 14 januari 2011. Het tv-programma, waarin deelnemers antwoorden moeten geven op allerlei vragen, is al sinds 1964 op de Amerikaanse televisie. Wat Watson uniek maakt(e) is de capaciteit om natuurlijke taal te interpreteren. In 1996 was er al de computer Beep Blue geweest die schaakkampioen Garry Kaparov versloeg, maar de '*problem space*', het schaakspel, is veel kleiner dan die waarop Watson werd ingezet.

Hoe werkt Watson?

Watson werkt volgens dezelfde vier principes als het menselijk denken bij het zoeken naar een antwoord op een gestelde vraag: observeren, interpreteren, evalueren en een keuze maken uit mogelijke antwoorden. Het observeren bestaat uit het raadplegen van allerlei gegevens uit verschillende bronnen, zoals wetenschappelijk literatuur, websites, blogs, enzovoort. Voor het grootste gedeelte zijn dit ongestructureerde gegevens in de vorm van vrije tekst. De kracht van Watson ligt in de zogenoemde *natural language processing*. Op dit moment kan dit alleen voor de Engelse taal. Aan Spaans en Portugees wordt gewerkt. De (super)computer wordt door menselijke experts gevoed met literatuur die relevant is voor een bepaald gebied. Op basis van vraag en antwoorden leert Watson. Zo ontstaat een zogenoemde *corpus of knowledge*. Op basis van deze kennis interpreteert en evalueert Watson observaties, zoekt naar verbanden en genereert hypothesen, in casu mogelijke antwoorden op een gestelde vraag. De verschillende hypothesen worden vervolgens getoetst aan de beschikbare informatie en gescored. De hypothese met de hoogste score vormt uiteindelijk het antwoord van Watson op een vraag. Watson evalueert continu nieuwe informatie die beschikbaar komt en breidt daarmee zijn *corpus of knowledge* uit. Zo ontstaat een zelflerend systeem.

Op dit moment wordt Watson ingezet voor het ondersteunen van artsen bij beslissingen over behandelingen. Hierbij wordt samengewerkt met een aantal medische faculteiten, zoals Columbia University, Memorial Sloan-Kettering Cancer Center in de Verenigde Staten en in Europa met universiteitsziekenhuizen in Graz en Wenen. Het uiteindelijke doel van IBM is om een digitale assistent, beschikbaar op een *mobile device*, zoals een *tablet* of *smartphone*, voor de arts gereed te hebben voor 2020⁴¹.

Bron: IBM⁴²

3.8 BETROUWBAARHEID VAN WETENSCHAPPELIJKE PUBLICATIES

Voor een adequate computerondersteuning van de zorg zijn naast een krachtige supercomputer(cluster), betrouwbare gegevens nodig. Deze gegevens zijn gebaseerd op de uitkomsten van wetenschappelijk onderzoek die gepubliceerd zijn in *peer-reviewed* wetenschappelijke tijdschriften.

Deze aanpak om te vertrouwen op wetenschappelijke literatuur is niet onproblematisch. In 2005 zorgde het artikel 'Why most published research findings are false' van Ioannides voor veel opschudding.⁴³ Hij liet zien dat de kans dat een onderzoeksresultaat juist is, afhankelijk is van een groot aantal factoren, waaronder de steekproefgrootte, mogelijke *bias* en het aantal andere studies op hetzelfde terrein. De kans op onjuistheid neemt toe naarmate een studie kleiner is, de effecten geringer zijn, er minder selectie is van de mogelijke verbanden en er dus meer worden onderzocht, de onderzoeksopzet minder rigide is en grotere flexibiliteit toestaat in ontwerp, gehanteerde definities, uitkomstmaten, er grotere financiële belangen op het spel staan en er meer onderzoeksgroepen bij betrokken zijn. Simulaties lieten zien dat gegeven de onderzoeksopzet van de meeste studies, de kans dat de onderzoeksresultaten onjuist zijn groter is dan de kans dat ze juist zijn.

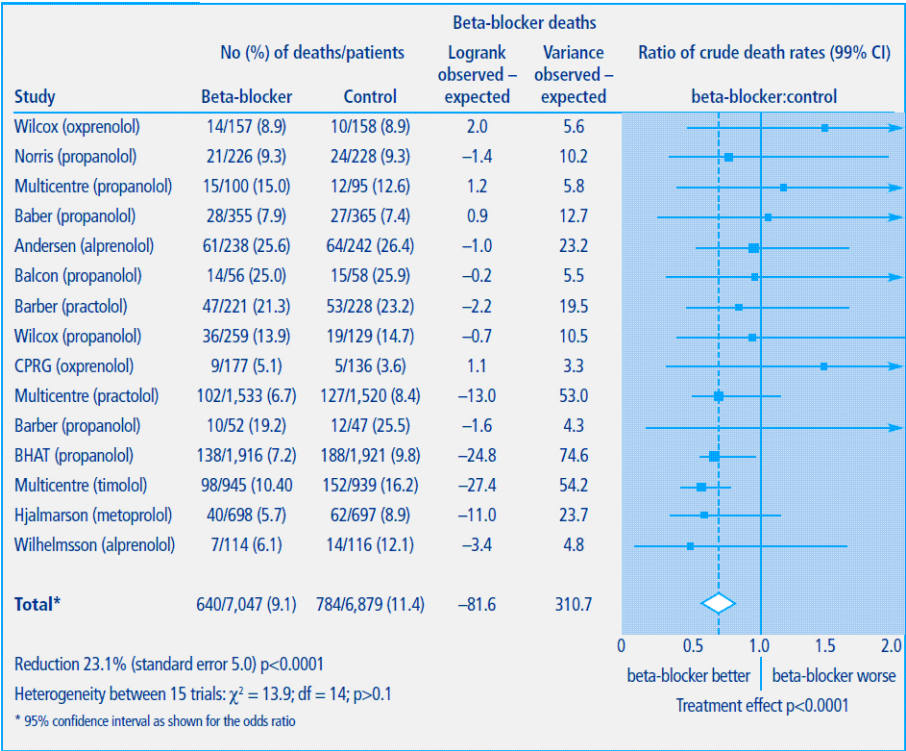
Een belangrijk instrument om het kaf van het koren te scheiden zijn de zogenoemde *Cochrane Reviews* door de Cochrane Collaboration, zie kadertekst. Het is niet uitzonderlijk dat bij een *Cochrane Reviews* het merendeel van de onderzoeken, tot wel 80%, in de prullenbak verdwijnt door een of meer van de hiervoor genoemde tekortkomingen in de onderzoeksopzet, zoals te kleine en/of niet aselecte steekproef.⁴⁴

Cochrane Review

Een *Cochrane Review* is een publicatie door de Cochrane collaboration. Dit is een internationaal netwerk van meer dan 28.000 wetenschappers in meer dan 100 landen, dat is opgericht in 1993. Het is genoemd naar Archie Cochrane (1909-1988), een Schotse arts en pionier op het gebied van *evidence-based medicine*. Doel is om belangrijke medische vragen, bijvoorbeeld naar de effectiviteit van een bepaalde behandeling, te beantwoorden op basis van wetenschappelijk bewijs.

Bij een *Cochrane Review* worden via literatuuronderzoek alle wetenschappelijke publicaties over een bepaald onderwerp verzameld en volgens een vast stramien geanalyseerd: de kwaliteit van de onderzoeksopzet, statistische bewerkingen, enzovoort. Het resultaat wordt meestal in de vorm van een zogenoemde *forest plot* weergegeven, zoals in het volgende voorbeeld⁴⁵.

Tabel 3.1Beta-blocker deaths

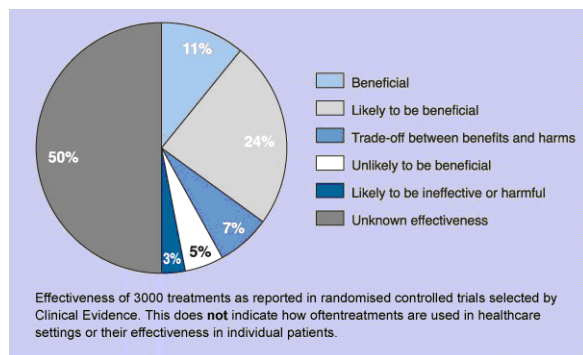


Bron: Cochrane Review⁴⁶

Elke studie wordt als een rondje of vierkantje weergegeven, waarbij de grootte een maat is voor de precisie van de studie. Hoe meer het rondje of vierkantje naar links ligt, hoe beter de behandeling, in dit voorbeeld met een betablokker (een groep geneesmiddelen die de bloeddruk verlagen), werkt c.q. lijkt te werken. Een horizontale lijn geeft het 95% betrouwbaarheidsinterval weer. De witte ruit geeft het gemiddelde van alle studies aan. In dit voorbeeld is het geaggregeerde behandelingsresultaat met een betablokker positief.

Alle *Cochrane Reviews* worden opgenomen in de *Cochrane Library*. Deze is in veel landen gratis toegankelijk, maar niet in Nederland.

Naast de vraag in hoeverre de beschikbare informatie betrouwbaar is, is een nog belangrijkere vraag over welke zaken überhaupt informatie beschikbaar is. In figuur 3.1 is aangegeven voor welk deel van veel voorkomende behandelingen er epidemiologische informatie beschikbaar is over de effectiviteit ervan. Het blijkt dat van de helft niet bekend is of deze behandelingen eigenlijk wel nut hebben, omdat er nooit goed onderzoek naar gedaan is. Er is een grote informatieleemte in de zorg.⁴⁷

Figuur 3.1 Effectiviteit van 3000 behandelingen

Bron: BMJ Clinical Evidence⁴⁸

3.9 KAN BIG DATA DE INFORMATIELEEMTE EN BETROUWBAARHEID VAN BESCHIKBARE GEGEVENS IN DE ZORG VERBETEREN?

In hoofdstuk 1, de inleiding, is aangegeven, dat de zorg weliswaar een grote hoeveelheid gegevens produceert, maar tegelijkertijd is er, zie de vorige paragraaf, een enorm gebrek aan informatie, bijvoorbeeld over de effectiviteit van behandelingen, en is veel van de informatie die wel beschikbaar is, onbetrouwbaar.

In paragraaf 3.7 is de supercomputer Watson besproken die onder andere de wetenschappelijk literatuur doorspint om zijn kennis aan te vullen. Maar wat nu als veel van die informatie onjuist blijkt te zijn? Op dit moment is er evenwel geen alternatief, maar de bestaande informatie zou aangevuld en op termijn wellicht vervangen kunnen worden door kennis die door de computer zelf is gegenereerd, zoals in paragraaf 3.5 is beschreven, op basis van Big Data die bijvoorbeeld aanwezig is in elektronische medische dossiers. Deze dossiers dienen daarbij wel aan een aantal eisen te voldoen. Zo moeten de gegevens van de patiënt nauwkeurig, volledig en gestandaardiseerd vastgelegd zijn. Helaas is dat op dit moment nog niet het geval. In het advies Patiënteninformatie heeft de RVZ (Raad voor de Volksgezondheid en Zorg) nadere aanbevelingen gedaan.⁴⁹ Dit kan aangevuld worden met gegevens die de patiënt zelf aanlevert, bijvoorbeeld via zijn of haar persoonlijk gezondheidsdossier. Een andere mogelijkheid is het gebruik van meetwaarden die direct op of in het lichaam worden gemeten en zonder tussenkomst van de patiënt (draadloos) worden doorgegeven (*machine-to-machine data*). Dit is onder andere het terrein van *Quantified Self*. Deze nieuwe gegevensbronnen komen in hoofdstuk 6 aan de orde. Een andere nieuwe bron van gegevens vormen de *high-throughput*-technieken die in het volgende hoofdstuk worden besproken.

3.10 CONCLUSIE

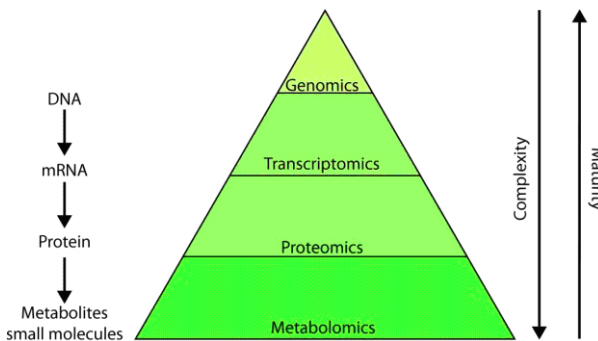
In dit hoofdstuk zijn de problemen bij de diagnosestelling en keuze voor behandelingsopties door de mens c.q. de arts beschreven, zoals *anchoring* en *confirmation bias* en het inschatten van kansen. Bijkomende problemen zijn dat de kennis van de oorzaken van ziekten en aandoeningen beperkt is en de huidige wetenschappelijke literatuur veel onjuistheden bevat. Door gebruik te maken van Big Data, bijvoorbeeld aanwezig in elektronische medische dossiers, kunnen kennislacunes opgevuld worden en de betrouwbaarheid vergroot. Een strikte kwaliteitsbewaking is hierbij van cruciaal belang, want ook voor Big Data geldt: *rubbish in rubbish out*.

4 HET ‘OMICS’-TIJDPERK

4.1 INLEIDING

De erfelijke informatie in een cel ligt opgeslagen in het DNA in de vorm van een sequentiële code. Deze code wordt gevormd door de volgorde van vier chemische bouwstenen: de basen adenine (A), thymidine (T), guanine (G) en cytosine (C). Vrijwel alle chemische reacties in de cel worden gestuurd door eiwitten en daarnaast vormen eiwitten belangrijke bouwstenen, niet alleen in de cel, maar ook daarbuiten. Eiwitten zijn opgebouwd uit een reeks aminozuren. Codes van drie basen in het DNA instrueren de eiwitfabriekjes in de cel, de ribosomen, welk aminozuur op een bepaalde plaats in de eiwitketen ingebouwd moet worden. Zo codeert bijvoorbeeld CAA voor het aminozuur valine en AAT voor leucine. Het transport van de informatie van DNA naar de ribosomen verloopt via RNA. Daartoe wordt een deel van het DNA dat codeert voor een eiwit – een gen – ‘overgeschreven’ naar RNA, de zogenoemde transcriptie. Zoals gezegd verzorgen eiwitten allerlei chemische omzettingen in de cel: de celstofwisseling of het metabolisme.

Figuur 4.1 Verschillende ‘omicsen’



Bron: BioMedCentral/Genome Medicine⁵⁰

Het ontrafelen van de genetische code van het DNA – het bepalen van de volgorde van de basen in het DNA, *sequencing*, was tot voor kort een moeizaam en zeer duur proces. Hetzelfde gold voor de bepaling van welke eiwitten en welke stofwisselingsproducten in een cel aanwezig zijn. Inmiddels is het mogelijk om met behulp van *zogenoemde high throughput technieken, waaronder next generation sequencing*, tegen redelijke kosten en binnen redelijke tijd de basenvolgorde van een groot deel van het DNA (genomics), van het RNA (transcriptomics), de eiwitten (proteomics) almede de stofwisselingsproducten (metabolomics) in een cel te bepalen. In figuur 4.1 is schematisch aangegeven dat genomics eigenlijk nog het ‘simpelst’ is: de basenvolgorde in het DNA bepalen en dat is het. De expressie van de genen via

RNA en eiwitten is veel complexer, waarbij in wisselwerking met de omgeving onder invloed van complexe regelmechanismen voor een in tijd en celtype variërend geheel van eiwitten en metabolieten ontstaan.

Met deze informatie tracht men te ontrafelen hoe processen in de cel werken, met name in interactie met de omgeving. Ziekten ontstaan uiteindelijk door ontregelingen in cellen, en inzicht in de chemische processen in de cel kan licht doen schijnen op de oorzaken ervan.

4.2 NEXT GENERATION SEQUENCING

Zoals hiervoor aangegeven is *sequencing* het bepalen van de volgorde van de basen in het DNA. Het hele genoom, dat in 23 paar chromosomen is vastgelegd, bestaat uit ruim 3 miljard basenparen. De eerste methoden voor de bepaling van de volgorde werden begin jaren zeventig van de vorige eeuw ontwikkeld. In 1977 ontwikkelde Sanger een methode waarmee de basenparen één voor één afgelezen konden worden.⁵¹ Voor lange tijd was deze methode, die nu *first generation sequencing* wordt genoemd, het meest gebruikt.

Inmiddels is het door de toepassing van nieuwe technologieën, waarbij computers een belangrijke rol spelen, mogelijk geworden om vele genen tegelijkertijd, parallel, te sequencen in één test. Men spreekt hierbij van *next generation sequencing*.⁵²

Het is belangrijk onderscheid te maken is verschillende soorten sequencing. Zo kan men gericht bepaalde genen sequencen (zogenoemde *target sequencing*). Een voorbeeld hiervan is de analyse van een tweetal genen BRCA1 en BRCA2, zie kader-tekst.

Screening op erfelijke borstkanker: BRCA1 en BRCA2

In 1990 ontdekte een groep onderzoekers aan de Universiteit van Californië een verband tussen een gen, BRCA1 genoemd, op chromosoom 17 en een verhoogd risico op borstkanker. In augustus 1994 hadden onderzoekers aan de Universiteit van Utah, de US Institutes of Health en de McGill University het gen gesequenced en er octrooi op aangevraagd. Dit octrooi werd in 1997 verleend. Vooruitlopend hierop had de firma Myriad Genetics in 1997 BRCAAnalysis op de markt gebracht. DNA kon naar het bedrijf opgestuurd worden, waarna zij bepaalden of BRCA1 aanwezig was of niet. In 1998 kregen dezelfde organisaties een octrooi op een tweede borstkankergen BRCA2, waarvoor Myriad ook een test aanbood.

De BRCA-genen kregen wereldwijde bekendheid toen op 14 mei 2013 de Amerikaanse actrice Angelina Jolie bekend maakte dat zij preventief haar borsten had laten amputeren, omdat zij drager bleek te zijn van een BRCA1-genmutatie met een kans van 87% op borstkanker.⁵³ Op 24 maart 2015 meldde zij dat zij ook haar eierstokken had laten verwijderen, vanwege een kans op kanker daarin vanwege dezelfde mutatie.⁵⁴

Er was veel kritiek op het monopolie van Myriad op basis van de octrooiering van menselijke genen. Behalve kritiek op de prijs, 4000 dollar, meer dan de kosten voor *whole exome sequencing* (zie tekst), was er vooral kritiek op het feit dat zij de informatie die zij vergaren met het analyseren van de *DNA-samples* voor zichzelf hielden. Vele rechtszaken volgden, met uiteindelijk de uitspraak van het US Supreme Court op 13 juni 2013 dat een natuurlijk voorkomend DNA-segment niet octrooieerbaar is, hetgeen de octrooiën van Myriad ongeldig maakte in de vs. Het bedrijf blijft evenwel aanbieders van testen met rechtszaken op andere gronden achtervolgen.⁵⁵

Een (grote) stap verder is het bepalen van de volgorde van alle basen in het DNA die voor eiwitten coderen. Het betreft hier de circa 22.000 genen die de mens bezit. Dit is 'slechts' 1% van de gehele genoom. Deze eiwitcoderende stukken worden *exon*-gen genoemd. Vandaar dat men in dit geval spreekt van *whole exome sequencing*.

Van de 99% niet voor eiwitten coderend DNA nam men tot voor kort aan dat dit nutteloze ballast was. Deze 'prullenbak van de evolutie' werd aangeduid met de term *junk DNA*. Inmiddels is men erachter gekomen dat deze niet-coderende gedeelten belangrijke functies vervullen bij de regulering van de genen. Bij *whole genome sequencing* wordt het gehele genoom gesequenced, dus zowel de genen als het niet voor eiwitten coderende gedeelte.

Door de ontwikkeling van de technologie wordt *whole genome sequencing* steeds goedkoper, naar verwachting op termijn zelfs goedkoper dan het sequencen van alleen (een aantal) genen.

4.3 SEQUENCING EN BIG DATA

Whole genome sequencing is zeker voor wat volume betreft Big Data bij uitstek. Een onbewerkt gegevensbestand van het genoom van één persoon beslaat ongeveer honderd Gigabyte. De huidige *sequence*-capaciteit in de wereld omvat zo'n 35 petabytes – dat zijn circa 250.000 humane genomen – per jaar.⁵⁶ Behalve volumineus is genetische informatie ook zeer complex. Verschillende technologieën worden gebruikt met verschillende accuraatheid en reproduceerbaarheid. Verschillende soorten weefsels kunnen om verschillende redenen gesequenced worden. Zo kan het 'geërfde genoom' van een persoon gesequenced worden, maar ook het gemuteerde genoom van een tumor. De resultaten kunnen verschillend gerapporteerd worden, bijvoorbeeld ziektespecifiek. Er kunnen echter ook incidentele of toevallsbevindingen zijn.⁵⁷ Daarnaast is er nog veel onbekend over de relatie tussen ziekte en bepaalde genmutaties (*variants of unknown significance*). Er is dan ook nog veel onderzoek c.q. *Big Data mining* nodig.

Ernstige erfelijke aandoeningen

Waar wel veel bekend is over de relatie tussen genen en ziekte is op het terrein van de, veelal ernstige, monogenetische erfelijke aandoeningen. Bij deze aandoeningen is er één gen gemuteerd en deze overerven volgens de wetten van Mendel. In de

databank Online Mendelian Inheritance waren in 2013 ruim 3700 aandoeningen beschreven waarvan de moleculaire basis opgehelderd is. Naar schatting 500 van deze aandoeningen veroorzaken reeds op kinderleeftijd ernstige symptomen. Een specifieke aandoening is zeldzaam. Zo komen stofwisselingsziekten, zoals de ziekte van Pompe, bij 14 op de 100.000 levend geboren voor. Per 100.000 geboortes komt *cystic fibrosis* komt bij 19, sikkelcelziekte bij 22 en thalassemie bij 14 levend geboren voor.⁵⁸

De ernstige erfelijke aandoeningen zijn per aandoening weliswaar zeldzaam, maar zoals hiervoor aangegeven zijn er ruim 3700 verschillende aandoeningen. Naar schatting leidt 0,25 tot 0,5% van de pasgeborenen aan een ernstige recessieve ziekte⁵⁹. Uit Amerikaans onderzoek blijkt dat iedereen gemiddeld drager is van 2,8 ernstige erfelijke aandoeningen.⁶⁰

Door dragerschap bij paren met kinderwens te bepalen, kan vóór de conceptie de kans op een ernstige erfelijke aandoening vastgesteld worden (preconceptioneel erfelijkheidsonderzoek). Indien er sprake is van een ‘match’ van een bepaalde recessieve aandoening, dan betekent dit dat het kind een kans van 25% op de aandoening heeft. In die situatie kan besloten worden tot pre-implantatie genetische diagnostiek. Met deze techniek kunnen genetische afwijkingen opgespoord worden in eicellen of embryo’s voorafgaand aan of aansluitend op in-vitrofertilisatie. Alleen niet-aangedane embryo’s worden in de baarmoeder teruggeplaatst. Opgemerkt dient te worden dat bij deze aanpak niets aan de genen wordt veranderd. De frequentie van dragerschap van ernstige erfelijke aandoeningen in de populatie verandert niet. De veronderstelling dat sprake zou zijn van eugenetica is dan ook ongegrond.

Op dit moment wordt preconceptioneel erfelijkheidsonderzoek met name aangeboden in gebieden waar bepaalde erfelijke aandoeningen relatief vaak voorkomen, zie kadertekst.

Preconceptiespreekuur Volendam

In Volendam komt een aantal ernstige erfelijke aandoeningen relatief vaak voor. Het gaat om pontocerebellaire hypoplasie type 2 (PCH2), osteogenesis imperfecta type II en III, rhizomele chondroplasia punctata type I en foetale akinesie.

De afdeling klinische genetica van het AMC is in 2012 een spreekuur gestart in Volendam. Paren met kinderwens kunnen zich laten onderzoeken op dragerschap van deze aandoeningen. Als beide partners drager van dezelfde aandoening blijken te zijn, dan hebben zij verschillende handelingsopties. Zo kunnen zij het risico accepteren, of kiezen voor prenatale of pre-implantatiediagnostiek, afzien van kinderen of adoptie overwegen. Na een jaar waren 196 personen onderzocht. Daarvan bleken 35 personen drager voor één of meer van de aandoeningen. Er werden vier paren gevonden met eenzelfde mutatie, namelijk voor PCH2). Deze paren hebben een kans van 25% op een kind met de betreffende aandoening.⁶¹

Met behulp van *next generation sequencing* kan dragerschap van in beginsel alle bekende ernstige erfelijke aandoeningen bepaald worden. Er is reeds een test beschikbaar voor het aantonen van dragerschap op 448 aandoeningen.⁶²

Begin 2015 kondigde de firma Illumina aan de 1000 dollar- grens voor het sequencen van het genoom doorbroken te hebben. Het ligt in de lijn der verwachting dat de kosten voor *next generation screening* zullen blijven dalen en binnen een aantal jaren binnen het bereik van grote groepen burgers zal komen. Deze genenpaspoorten zullen een enorme hoeveelheid Big Data opleveren. In eerste instantie was het idee dat hiermee de gevoeligheid voor ziekten, zoals hart-en vaatziekten, in kaart kon worden gebracht. Dit blijkt lastiger dan in eerste instantie gedacht, maar voor preconceptioneel onderzoek en de hierna te bespreken farmacogenetica is het op dit moment reeds toepasbaar.

4.4 FARMACOGENETICA

Geneesmiddelen worden vooral in de lever, het eerste orgaan dat een geneesmiddel tegenkomt na opname in de darm, omgezet in inactieve of juist actieve stoffen. Hierbij spelen de zogenoemde *cytochrome p450*-eiwitten een belangrijke rol. De mens heeft meer dan 60 genen, onderverdeeld in 14 subfamilies, die voor deze eiwitten coderen. Deze dragen namen als CYP1A2, CYP1B1, CYP2C9 enz.

Er bestaan allerlei genetische verschillen tussen mensen in hun vermogen om geneesmiddelen om te zetten, die veelal terug te voeren zijn op mutaties in de cytochrome p450 genen. Sommigen breken een geneesmiddel sneller of juist minder snel af dan de meeste andere mensen. Men spreekt dan van snelle resp. trage metaboliseerders. Als iemand een snelle metaboliseerder is, dan levert een normale dosis van het geneesmiddel een lagere concentratie in het bloed op, waardoor de werkzaamheid afneemt. Bij een trage metaboliseerder kan een normale dosis zich ophopen waardoor gevaarlijke bijwerkingen kunnen optreden.

De genetische verschillen betreffen meestal de wijziging van één nucleotide in een gen, zogenoemde *single nucleotide polymorphismen* (SNP's).

Er zijn inmiddels verschillende tests beschikbaar voor een aantal geneesmiddelen, zoals CYP2C9 en VKORC1 voor antistollingsmiddelen zoals fenprocoumon. In tabel 4.1 is als voorbeeld een aantal genen waarop getest kan worden voor geneesmiddelen waarvan de dosis mogelijk aangepast moet worden, weergegeven.

Tabel 4.1 Een aantal geneesmiddelen en bijbehorende genetische tests

fenprocoumon	CYP2C9, VKORC1	tacrine	CYP1 A2
fentanyl	CYP3A4	tacrolimus	CYP3A4, CYP3A5, POR
fentytoïne	CYP2C9, CYP2C1 9	tamoxifen	CYP2D6*
finasteride	CYP3A4	tegafur	DPYD
flecaïnide	CYP2D6	temsirolimus	CYP3A4
flucloxaciline	HLA-B*5701 (bijwerkingen)	teniposide	CYP2C1 9
fluorouracil (5-FU)	DPYD	thioguanine	TPMT
fluoxetine	CYP2C9, CYP2D6	thioridazine	CYP2D6
fluvastatine	CYP2C9, CYP3A4	timolol	CYP2D6
fluvoxamine	CYP1 A2, CYP2D6	tizanidine	CYP1 A2
glibenclamide	CYP2C9	tolbutamide	CYP2C9
glimepiride	CYP2C9	tramadol	CYP2D6
glipizide	CYP2C9	triazolam	CYP3A4
haloperidol	CYP2D6, CYP3A4	venlafaxine	CYP2D6, CYP3A4
halothaan	CYP2E1	verapamil	CYP1 A2
hexobarbital	CYP2C1 9	vincristine	CYP3A4, CYP3A5
ibuprofen	CYP2C9	voriconazol	CYP2C9, CYP2C1 9
ifosfamide	CYP2B6, CYP3A4*	warfarine	CYP2C9, VKORC1
imatinib	CYP3A4	ziprasidon	CYP3A4
imipramine	CYP1 A2, CYP2C1 9, CYP2D6	zolmitriptan	CYP1 A2
indinavir	CYP3A4	zolpidem	CYP3A4
indometacine	CYP2C9	zuclopentixol	CYP2D6

*=activatie, m=metaboliët, rood=dosisadvies KNMP, groen=geen aangepaste dosis nodig volgens KNMP, zwart=niet beoordeeld door KNMP. Bovenstaande informatie dient alleen als hulpmiddel. Er kunnen geen rechten aan worden ontleend.

Bron: ErasmusMC

Met het beschikbaar komen van de testen is het mogelijk geworden om dosering van geneesmiddelen te personaliseren, dat wil zeggen aan te passen aan de genetische kenmerken van de patiënt. Hierdoor kan over- of onderdosering, met soms fatale gevolgen, voorkomen worden. Deze farmacogenetica is een vorm van *personalized medicine*, die in paragraaf 5.3 nader besproken wordt.

4.5 TRANSCRIPTOMICS, PROTEOMICS, METABOLOMICS

Op het terrein van *transcriptomics*, *proteomics* en *metabolomics* zijn soortgelijke technologische ontwikkelingen gaande als bij *genomics*. Steeds grotere aantallen kunnen steeds sneller geanalyseerd worden tegen lagere kosten. Vooral nog vinden deze bepalingen voornamelijk plaats in een onderzoekscontext. Het relateren van de grote hoeveelheden gegevens aan elkaar is het terrein van de systeembiologie, die in het volgende hoofdstuk wordt besproken.

4.6 AANDACHTSPUNTEN BIJ DE NIEUWE TECHNOLOGISCHE ONTWIKKELINGEN

In zijn advies Next generation sequencing noemt de Gezondheidsraad een aantal aandachtspunten. Behalve een aantal meer technische zaken, zoals de kwaliteit van de *sequencing* en de analyse en interpretatie, betreffen het vooral ethische en juridische aspecten.⁶³ Een belangrijk aandachtspunt daarbij is de privacy. Wie mag de gegevens beheeren? Hoe lang mogen/moeten deze bewaard blijven? Wie bepaalt dat? Wie mag de gegevens inzien? Wie houdt toezicht? De Gezondheidsraad noemt Estland als voortrekker op het terrein van de privacy. De gezondheidszorg en maatschappelijke dienstverlening zijn daar verregaand gedigitaliseerd. De individuele burger kan zien wie welke gegevens heeft ingezien en kan bezwaar aantekenen als dit onrechtmatig is gebeurd. Daarnaast kan hij of zij bepaalde gegevens blokkeren, zie kadertekst.

Estland

Toen Estland in 1991 onafhankelijk werd van de Sovjet-Unie, stond het land er niet zo goed voor. De (zorg)voorzieningen en financiële middelen waren beperkt. Dit was evenwel ook een voordeel. Het was het begin van het internettijdperk en de regering zette hierop in. Een infrastructuur werd opgezet, met een nationale elektronische identiteitskaart en een elektronisch medisch dossier. De ID-kaart is beveiligd met een 2048-bit *public-key* versleuteling en wordt onder ander gebruikt als nationale zorgverzekeringskaart en toegang tot medische gegevens, zoals het elektronisch medisch dossier. Artsen kunnen onderzoeksresultaten, waaronder röntgenfoto's enzovoort raadplegen. Patiënten hebben met hun ID-kaart via een patiëntenportaal toegang tot hun medische gegevens en dat van hun kinderen. Zij kunnen aangeven welke zorgverleners toegang hebben tot hun dossier. De medische gegevens worden gebruikt voor statistische doeleinden, zoals het monitoren van de volksgezondheid en de doelmatigheid van de zorgverlening en het volgen van (griep)epidemieën.

Niet alle initiatieven zijn evenwel een succes. Zo werd in 2000 het Estlands *genome project* in het leven geroepen. Het plan was om binnen 10 jaar van 70% van de 1,4 miljoen Esten biomateriaal en gezondheidsgegevens te verzamelen. In 2014 werd dit doel bijgesteld naar 5%. Tot dan toe was slechts van 52.000 mensen biomateriaal verzameld.

Bron: e-Estonia

De goede toegankelijkheid van medische gegevens van de patiënt via de nationale elektronische ID-kaart door behandelend artsen en de patiënt zelf dragen bij aan de kwaliteit van de zorgverlening. De universele ID-kaart maakt het systeem evenwel ook kwetsbaar. De kaart en de elektronische gegevensuitwisseling zijn weliswaar goed beveiligd, maar als het kwaadwillenden lukt om de kaart te kraken, liggen de medische gegevens van alle Estlanders op straat.

In het Estlandse voorbeeld wordt ook het verzamelen van biomateriaal genoemd. Dit biomateriaal wordt opgeslagen in biobanken en biedt ongekekende mogelijkheden voor wetenschappelijk onderzoek. Maar ook hier is de vraag of de privacy voldoende gewaarborgd kan worden, aangezien het biomateriaal zelf de informatie bevat voor het identificeren van een individu.

4.7 VERVAGENDE GRENZEN TUSSEN DIAGNOSTIEK EN ONDERZOEK

Andere door de Gezondheidsraad gesignaleerde problemen zijn de vervagende grenzen tussen diagnostiek en onderzoek en diagnostiek en screening. Indien er bij verdenking op een genetische oorzaak van gezondheidsproblemen bij een patiënt ongericht wordt gesequenced, is er dan sprake van diagnostiek of van wetenschappelijk onderzoek? Zoals hiervoor aangegeven is er nog veel onbekend over de relatie tussen genen en ziekte, maar deze kennis zal snel toenemen. De vraag rijst wanneer patiënten geïnformeerd moeten of mogen worden – *recontacting* – indien er bijvoorbeeld nieuwe informatie beschikbaar komt over een bij hen gevonden genmutatie. Bij een onderzoek kunnen ook risico's op andere aandoeningen gevonden worden, waardoor de grens tussen diagnostiek en screening vervaagt. Hiermee nauw verbonden is het *informed consent*. Patiënten moeten geïnformeerde toestemming geven voor een medisch onderzoek, dus ook voor genetisch onderzoek. Maar waar geeft de patiënt dan precies toestemming voor? *Whole genome sequencing* kan immers duizenden risicofactoren voor aandoeningen opleveren. Sommige daarvan zijn te behandelen of daarvan is het risico door bijvoorbeeld *lifestyle* aanpassingen te verkleinen, maar een aantal niet. Welke uitslagen wil de patiënt wel weten en welke niet?

4.8 CONCLUSIE

Op dit moment zijn er stormachtige ontwikkelingen gaande op het terrein van de 'omics': genomics, transcriptomics, proteomics en metabolomics. Met de komst van *high-throughput*-technieken zoals *next generation sequencing* neemt de kennis van de genetische component van aandoeningen enorm toe. Door de sterk dalende kosten van de technieken is ook de mogelijkheid ontstaan om de genetische constitutie van de individuele patiënt in kaart te brengen, wat onder andere farmacogenetica en preconceptionele screening mogelijk maakt.

De toegenomen mogelijkheden roepen evenwel nieuwe vragen op, met name op het terrein van de ethiek en privacy: wie mag de gegevens beheren, wie mag ze inzien, wie houdt toezicht, is de privacy gewaarborgd, enzovoort. In hoofdstuk 9 wordt op deze aspecten nader ingegaan.

5 SYSTEEMBIOLOGIE, SYSTEMS MEDICINE EN P4 MEDICINE

5.1 INLEIDING

De traditionele wetenschappelijke aanpak is om een complex systeem uit elkaar te halen en kennis te verzamelen over onderdelen ervan. Een voorbeeld hiervan is de *sequencing* van het DNA zoals in het vorige hoofdstuk beschreven. Deze reductio-nistische aanpak is zeer succesvol voor het begrijpen van hoe een deelsysteem werkt. Het heeft ook grote successen in de geneeskunde opgeleverd. Zo kunnen bijvoorbeeld door *sequencing* van het DNA genmutaties opgespoord worden die ernstige erfelijke aandoeningen veroorzaken. Doordat bekend is welk eiwit defect is, kan gericht gezocht worden naar een behandeling. Bij veel met name chronische aandoeningen, zoals diabetes, kanker en auto-immuunziekten, ligt het veel complexer en moet gekeken worden naar het geheel. Dit is het terrein van de systeem-geneeskunde, ook vaak netwerkgeneeskunde genoemd en meestal aangeduid met de Engelse term *systems/network medicine*. Een stap verder is om op basis hiervan op de individuele patiënt toegesneden behandelingen te ontwikkelen. Dit wordt aangeduid met de term *P4-medicine*. Bij al deze ontwikkelingen speelt Big Data een cruciale rol, zoals in dit hoofdstuk aan de orde komt.

5.2 SYSTEMS MEDICINE

Systeem- of netwerkbioogie

Biologische systemen zoals het menselijk lichaam zijn zeer complex. Zo is het lichaam opgebouwd uit organen, zoals hersenen, lever, darm, spieren en huid, die elk weer zijn opgebouwd uit vele cellen. Al deze cellen (van één individu) bevatten in beginsel dezelfde erfelijke code, hetzelfde DNA, maar toch is een spiercel anders dan een levercel. Al deze cellen in de verschillende organen werken met elkaar samen, zodat we als (gezond) mens kunnen functioneren. De systeem- of netwerkbioogie kijkt naar het leven als één geheel van biologische, hiërarchisch georganiseerde netwerken waar organen, cellen, eiwitten en genen elkaar voortdurend beïnvloeden in wisselwerking met de omgeving. Het ultieme doel is netwerken in kaart te brengen en inzicht te krijgen in hoe deze met elkaar verbonden zijn.⁶⁴

Reductionistische geneeskunde

Zoals in de inleiding is aangegeven zijn er ziekten met een enkelvoudige oorzaak, zoals de in het vorige hoofdstuk beschreven veelal ernstige, aangeboren erfelijke aandoeningen waarbij één gen niet (goed) werkt. Verreweg de meeste ziekten, en met name de chronische aandoeningen, zoals kanker en hart- en vaatziekten, zijn echter complex en zijn het resultaat van het samenkomen van verschillende facto-

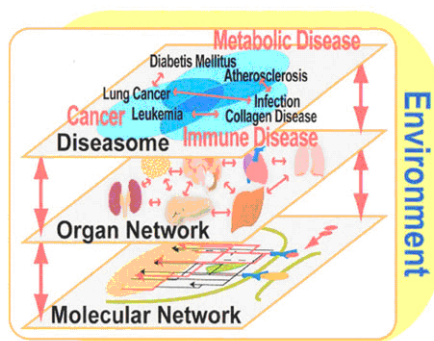
ren. Dit kunnen naast erfelijke ook allerlei omgevingsfactoren zijn. Daarnaast kunnen met name oudere patiënten meerdere aandoeningen tegelijkertijd hebben, die elkaar onderling beïnvloeden.

De geneeskunde is echter reductionistisch georganiseerd op basis van orgaanspecialisatie, bijvoorbeeld cardiologie, neurologie, oogheelkunde. Behandeling van een complex gezondheidsprobleem vindt vaak plaats vanuit los van elkaar opererende disciplines. Zo komt het veelvuldig voor dat patiënten door verschillende medisch specialisten geneesmiddelen voorgeschreven krijgen die elkaar kunnen tegenwerken. Van de 65-plussers in Nederland gebruikt naar schatting 30-40%, ca. 750.000 – 1 miljoen mensen, dagelijks vijf of meer verschillende geneesmiddelen.⁶⁵ Deze polyfarmacie kan op zichzelf weer tot ziekte (en ziekenhuisopnames) leiden.⁶⁶

Systems Medicine/Network Medicine

Aangezien complexe gezondheidsproblemen betrekking hebben op de gehele patiënt en zich uitstrekken over verschillende orgaansystemen tracht *systems/network medicine* de systeem/netwerkaanpak toe te passen binnen de zorg. In onderstaand figuur is dit schematisch weergegeven.

Figuur 5.1 Systeem/netwerkaanpak binnen de zorg



Bron: Tohoku University

In het vorige hoofdstuk is aangegeven dat de hoeveelheid en diversiteit van de gegevens op het 'omics'-terrein – het moleculaire netwerkniveau in de figuur – enorm is. Dit koppelen aan het orgaannetwerkniveau en aan gezondheid en ziekte – het *diseaseome* in de figuur – kan gezien worden als 'Big Data tot de derde macht'. *Systems Medicine* is dan ook ondenkbaar zonder wiskunde, statistiek en computers. Het vraagt ook een nauwe samenwerking tussen biologen, chemici, natuurkundigen, wiskundigen, informatici en medici.⁶⁷

5.3 P4 MEDICINE

Binnen het concept van P4 Medicine staat P4 voor de vier P's van *predictive, preventive, personalized* en *participatory*. De term is bedacht door Hood, mede-oprichter en president van het Institute for Systems Biology in Seattle.

Predictive, preventive, personalized

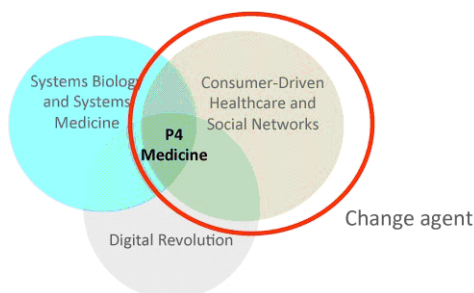
Op basis van systeembioologische gegevens van een individu, zoals zijn of haar genoom, proteoom, metaboolom en microbiom, kunnen individuele verschillen meegenomen worden bij de preventie of behandeling van ziekten: *personalized medicine*.

Het idee van Hood is om van iedereen het genoom en proteoom te bepalen. Daarnaast zou regelmatig de functie van 50 organen gecheckt moeten worden door per orgaan de bloedwaarden van 50 eiwitten te meten. Per individu betekent dit 2500 meetwaarden. Door de metingen regelmatig te herhalen, bijvoorbeeld in de vorm van een jaarlijkse check-up, kan het verloop van de gezondheidstoestand bepaald worden. Bij afwijkingen kan preventief, in een vroeg stadium, nog voordat er symptomen zijn opgetreden en een ziekte zich heeft geopenbaard, ingegrepen worden.⁶⁸

Op dit moment is de aanpak van Hood nog niet haalbaar, maar de richting van *personalized medicine* is reeds ingeslagen. De eerder besproken farmacogenetica is hier een voorbeeld van. Een ander voorbeeld is de behandeling van kanker waarbij op basis van tumorkenmerken bepaald wordt welke (combinatie van) medicijnen voor de betreffende patiënt het meest effectief zijn.

Participatory

De participatoire component van P4 vloeit voort uit 'de digitale revolutie', waarbij netwerken van consumenten/patiënten als belangrijke drijvende kracht worden gezien voor veranderingen in de zorg, zie de *Change Agent* in figuur 5.2. De consument/patiënt wordt van passieve ontvanger van advies van een expert c.q. de zorgverlener tot een deelnemer in actieve sociale netwerken gericht op het bevorderen en behouden van gezondheid en het zelf managen van ziekte. In figuur 5.2 zijn de drie ontwikkelingen, systeembioologie/*Systems Medicine*, de digitale revolutie en door consumenten gestuurde zorg en sociale netwerken weergegeven. P4 Medicine bevindt zich op het snijvlak van deze drie.

Figuur 5.2 P4 Medicine met de patiënt als stuwende kracht

Bron: P4-medicine instituut

P4-medicine is bij uitstek een voorbeeld van Big Data in de zorg : enorme hoeveelheden data van verschillende aard en uit verschillende bronnen komen samen met als doel tot de beste, op het individu toegesneden behandeling te komen.

5.4 SYSTEMS MEDICINE EN P4 MEDICINE EN OVERHEIDSBELEID

Systeembioogie was een centraal thema in de NWO-strategienota 2007-2010 en vormde de aanzet voor een strategisch actieplan. Daarbij werden drie terreinen onderscheiden: de *rode* systeembioogie, die betrekking heeft op de medische wetenschappen, de *groene*, gericht op voeding (agro-food) en de *witte*, gericht op de industriële biotechnologie. Inmiddels zijn drie centra voor rode systeembioogie opgericht die worden ondersteund door NWO en ZonMw. Het betreffen het Nijmeegse Centre for Systems Biology and Bioenergetics (CSBB), het Groningse Systems Biology Center for Energy Metabolism and Aging (SBC-EMA) en het Amsterdamse Cancer Systems Biology Center (CSBC). In april 2015 is de Systems Medicine Web Hub online gegaan. Het stimuleert systeembenaderingen in de geneeskunde en ondersteunt onderzoekers, projecten en initiatieven met communicatie over nieuws, publicaties, bijeenkomsten, enzovoort.⁶⁹

Ook op Europees niveau wordt het belang van *Systems Medicine* gezien. Verschillende landen investeren in het opzetten van interdisciplinaire onderzoekscentra en de Europese Commissie stelde 360 miljoen euro beschikbaar binnen het Europese onderzoeksprogramma (het 7e Kaderprogramma dat liep van 2007 tot 2013) voor fundamenteel onderzoek.

5.5 CONCLUSIE

In dit hoofdstuk zijn de mogelijkheden (en problemen) van *high-throughput*-technieken geschetst. De integratie van grote hoeveelheden gegevens die betrekking hebben op verschillende niveaus: van molecuul, cel, orgaan tot individuen en

omgeving – *Systems Medicine* – maakt P4 Medicine mogelijk: predictieve, preventieve, gepersonaliseerde en participatoire zorg. Het doel is afwijkingen in de gezondheidstoestand van een individu op te sporen en preventief in te grijpen, dus nog voordat een ziekte zich heeft geopenbaard. Binnen dit concept is een centrale rol weggelegd voor (netwerken van) consumenten/patiënten. Zij worden als een belangrijke drijvende kracht gezien voor veranderingen in de zorg: de participatoire component in de vorm van *consumer-driven healthcare* en sociale netwerken. Hier zijn instrumenten voor nodig, van de basale internetverbindingen, sociale netwerksites tot de *apps* en apparaten om als patiënt allerlei zaken te meten. Deze worden door het bedrijfsleven ontwikkeld. Dit komt in het volgende hoofdstuk aan de orde.

6 DE PRIVATE SECTOR EN BIG DATA IN DE ZORG

6.1 INLEIDING

Het internet is ontstaan vanuit een initiatief van de Amerikaanse overheid vanuit strategisch militair oogpunt, dat resulteerde in DARPA-net. Daarna ontwikkelde het zich in de academische wereld. Door toedoen van private partijen werd het internet uiteindelijk gemeengoed, alhoewel ook hier een niet-private partij bij betrokken was, namelijk Tim Berners-Lee, werkzaam bij CERN, die het wereldwijde web bedacht, een cruciale rol speelde. In hoofdstuk 2 is beschreven hoe de technische beperkingen waar zoekmachine Yahoo! tegenaan liep mede de basis legde voor de ontwikkeling van Big Data. Inmiddels is, als het om het ontsluiten van het web gaat, het stokje overgenomen door Google.

Het ontsluiten van informatie is van oudsher een taak van (collectief gefinancierde) bibliotheken. Op het internet is het evenwel een vrijwel geheel private gelegenheid geworden. Op het terrein van medisch wetenschappelijke artikelen is er nog wel publieke ontsluiting in de vorm van Medline. Deze databank wordt onderhouden door het Amerikaanse National Network of Libraries of Medicine. Er is ook een commerciële tegenhanger in de vorm van Google Scholar. Big Data is een zaak geworden van grote bedrijven zoals Google, Apple, Facebook, Twitter, Samsung, Microsoft en IBM. In hoofdstuk 3 zijn de activiteiten van laatstgenoemd bedrijf rond de supercomputer Watson reeds beschreven. Deze bedrijven hebben inmiddels de zorg als potentiële groeimarkt ontdekt. Een belangrijke aanjager hiervan zijn de zogenoemde *lifestylegadgets*, die hun oorsprong vinden in de *quantified-self* ontwikkeling.

6.2 QUANTIFIED SELF

De term *quantified self* is bedacht in 2007 door Gary Wolf en Kevin Kelly, redacteuren van Wired Magazine.⁷⁰ De ontwikkeling is evenwel al in de jaren zeventig van de vorige eeuw gestart, waarbij draagbare sensoren in combinatie met draagbare computers en draadloze communicatie werden gebruikt.⁷¹ Het verzamelen van gegevens beperkt zich niet tot fysiologische c.q. gezondheidsgegevens, het kunnen ook gegevens betreffende bijvoorbeeld dagbesteding zijn. De gegevens kunnen ook verzameld worden bij baby's: *quantified baby*. Hierbij worden allerlei gegevens over de dagelijkse activiteiten geregistreerd, zoals voedings- en slaaptijden, tijdstip van verschonen, enzovoort. In de Verenigde Staten zijn vooral zogenoemde *sleep trackers* voor baby's populair, omdat ouders daar bevreesd zijn voor wiegendood. De effectiviteit van deze *baby monitors* is overigens niet aangetoond.⁷²

Zoals gezegd bestaat *quantified self* reeds een aantal decennia, maar de afgelopen jaren is het in een stroomversnelling gekomen door de beschikbaarheid van kleine en goedkope sensoren en computers met draadloze communicatie in de vorm van *smartphones*. Het gaat vooral om *lifestyle gadgets*, bijvoorbeeld fitnessarmbanden, horloges en weegschalen, waarmee fysiologische zaken, zoals lichaamsbeweging, slaap(bewegingen) en hartfrequentie gemeten kunnen worden. Daarnaast zijn er nog draadloze bloeddruk- en zuurstofsaturatiemeters. De meetgegevens worden meestal draadloos doorgegeven aan een *smartphone*, die de gegevens, eveneens draadloos, via het internet verder kan versturen. Het ligt in de lijn der verwachting dat er in de toekomst nog veel meer zaken met kleine, relatief goedkope *consumenten-ehealth*-apparaatjes gemeten kunnen worden. Hierbij kan gedacht worden aan ECG-monitoring, elektronische stethoscopen, oor- en oogspiegels en longfunctiemeters.⁷³ Wellicht komt in de toekomst zelfs echoscopie binnen het bereik van de consument. Er bestaat voor de professionele markt reeds een echoscopie-apparaat bestaande uit een kleine *ultrasound transducer* verbonden met een *smartphone*.⁷⁴

In het geval dat de apparaatjes automatisch de meetwaarden via het internet doorsturen naar een andere computer, bijvoorbeeld bij monitoring van de bloeddruk, hartfrequentie of het ECG, vormen ze een onderdeel van het *Internet of Things* (IoT). Deze term is bedacht door de Engelse ondernemer Kevin Ashton in 1999. Het IoT is veel breder dan de zorg.⁷⁵ Het gaat om allerlei apparaten die via een netwerk, zoals het internet, met elkaar verbonden zijn. In paragraaf 7.7 worden een tweetal voorbeelden gegeven, namelijk het meten van fijnstof en van radioactiviteit met de *smartphone*.

6.3 HEALTH-PLATFORMS

Een probleem bij het gebruik van de hiervoor besproken *lifestyle gadgets* van verschillende leveranciers is dat er geen integratie van de gegenereerde gegevens mogelijk is. Voor elk apparaatje moet een andere *app* of website geraadpleegd worden.

Om dit probleem (deels) op te lossen hebben bedrijven zoals Apple, Google en Samsung elk hun eigen zogenoemde *health*-platform ontwikkeld. Een dergelijk platform levert ontwikkelaars van *apps* gereedschappen, zodat de informatie van verschillende *apps* geïntegreerd aan de gebruiker getoond kan worden, althans binnen een specifiek platform. Het platform van Apple heet *HealthKit*. Vanaf versie 8 van het besturingssysteem *ios8* voor de *iphone* en *ipad*, is de *app* Gezondheid standaard aanwezig. Hiermee kunnen de gegevens binnen de *HealthKit*-omgeving getoond worden. Gebruikers kunnen zelf aangeven welke gegevens van welke sensoren c.q. apparaten, zoals de Apple Watch, zij met de *app* willen delen.

Google heeft zijn eigen platform voor het Android besturingssysteem in de vorm van *Google Fit*. Verschillende leveranciers maken fitnessgadgets, zoals *Android Wear smartwatches*, die gegevens aanleveren. Bij Samsung heet het platform SAMI. Aangezien Samsung geen eigen besturingssysteem heeft – de *smartphones* werken met Android – zoekt deze de oplossing in de cloud. Gezondheidsgegevens die verschillende apparaatjes produceren worden opgeslagen in de *cloud* en daar geanalyseerd en aan de gebruiker getoond.

De verschillende leveranciers lijken verschillende strategieën te hanteren. Zo richt Google zich vooralsnog specifiek op *lifestyle*. In het verleden was er het project Google Health waarin alle medische gegevens die op verschillende plaatsen door verschillende zorgverleners werden bijgehouden, bijeengebracht werden. Het project sloeg echter niet aan en werd in 2013 gestaakt. Apple daarentegen richt zich (ook) op de zorg. Zo werkt zij samen met bijvoorbeeld de Mayo Clinic in de VS en met ziekenhuisinformatiesysteemleverancier EPIC.

Platform voor medisch onderzoek

Bij de presentatie van de Apple Watch op 9 maart 2015 kondigde de CEO van Apple een nieuw *open source* platform aan onder de naam *ResearchKit*. Gegevens van de sensoren van het horloge, zoals hartslagmeter en stappenteller, die in de Gezondheid-app worden opgeslagen, kunnen met de *ResearchKit* door de gebruiker beschikbaar gesteld worden aan de medische wetenschap. Een aantal onderzoekscentra, zoals Stanford en Oxford hebben hun naam reeds aan het nieuwe platform verbonden. Naast gegevens van sensoren kunnen ook gegevens van *apps* gedeeld worden. Zo kan met een *app* de reactiesnelheid van de gebruiker gemeten worden. Daarnaast kan de gebruiker vragenlijsten op de *smartphone* invullen.

Tegelijk met de aankondiging van de *ResearchKit* werden de eerste vijf gezondheidsapps die met dit platform gebouwd waren, getoond. Die richten zich op een vijftal aandoeningen: Parkinson, diabetes, cardiovasculaire ziekten, COPD en borstkanker. Met de Parkinson-app, genaamd *mPower*, kan ondermeer via de microfoon van de *iphone* de stem opgenomen worden. Deze kan dan geanalyseerd worden op trillingen. Via de versnellings sensoren in de *iphone* kunnen loopbewegingen geanalyseerd worden.⁷⁶ Bij de *app* voor onderzoek naar cardiovasculaire aandoeningen, genaamd *MyHeart Counts*, wordt onder andere de hartslag gemeten gedurende een 6-minuten-looptest. Daarnaast wordt de lichamelijke activiteit gevolgd en worden via een vragenlijst de risicofactoren voor hart- en vaatziekten in kaart gebracht.⁷⁷ Een dag na de aankondiging hadden reeds 11.000 mensen zich ingeschreven voor de studie naar cardiovasculaire aandoeningen door Stanford Cardiovasculair Health.⁷⁸

6.4 MEDISCH ONDERZOEK EN PRIVATE FINANCIERING

Door toepassingen zoals de *ResearchKit* kunnen snel grote aantallen patiënten betrokken worden bij medisch onderzoek. Het vormt een concrete invulling van participatoire zorg, niet alleen bij de eigen behandeling, maar ook voor het genereren van nieuwe wetenschappelijke kennis.

In paragraaf 3.8 is aangegeven dat er grote lacunes in kennis bestaan in de zorg, bijvoorbeeld met betrekking tot de effectiviteit van behandelingen. Big Data-toepassingen zoals de *ResearchKit* kunnen een belangrijke bijdrage bieden aan het opvullen hiervan.

In dezelfde paragraaf is het onderzoek van Ioannides beschreven ten aanzien van de (huidige) kwaliteit van wetenschappelijke onderzoeksuitkomsten. Hij geeft aan dat naarmate er grotere financiële belangen op het spel staan, de kans op onderzoeks*bias* toeneemt. Daarnaast sturen financiële belangen ook wat er onderzocht wordt. Zo worden bijvoorbeeld veel grote klinische studies in opdracht van de farmaceutische industrie verricht. Het betreft onderzoek omtrent geotrooieerde geneesmiddelen. Zodra een geneesmiddel uit patent is, is onderzoek ernaar voor fabrikanten niet meer interessant.

Big Data-toepassingen in de zorg, zoals de *ResearchKit*, moeten op één of andere wijze gefinancierd worden. Wie betaalt, bepaalt. Een belangrijke vraag is dan ook, hoe dit in de praktijk invulling gaat krijgen. Indien de financiering grotendeels vanuit de private sector plaatsvindt, zal deze ook een grote invloed hebben op de onderzoeksagenda. De belangen van de private sector behoeven niet parallel te lopen met die van de patiënt of in bredere zin met de volksgezondheid, met een reële kans dat de lacunes in kennis selectief en gebiased worden opgevuld.

6.5 CONCLUSIE

Big Data-ontwikkelingen zoals consumenten-*eHealth* en wetenschappelijk onderzoek op brede schaal via internet via *frameworks* zoals de *ResearchKit* van Apple bieden ongekennde mogelijkheden voor het genereren van gegevens waarmee de grote lacunes in kennis in de zorg, zoals beschreven in hoofdstuk 3, opgevuld kunnen worden. Een belangrijk vraagstuk hierbij is hoe dit vorm gaat krijgen. Als de gegevensstromen via private partijen lopen, zal er commerciële invloed zijn op de richting en wijze waarop de kennislacunes worden opgevuld.

7 DE OMGEVING: EPIDEMIOLOGIE EN BIG DATA

7.1 INLEIDING

In figuur 5.1 is de omgeving – *environment* – in de achtergrond aangegeven alsof zij niet zo relevant is, maar niets is minder waar. Verreweg de meeste, zo niet alle, ziekten ontstaan in wisselwerking met de omgeving. Zaken als voeding en *lifestyle* hebben grote invloed.

Een omgeving die de mens letterlijk ‘op de huid’ zit, is de microflora op het lichaam. Het omvat de microbenpopulaties, zoals virussen, bacteriën, gisten en schimmels die zich op de diverse plaatsen op de huid en in de darm bevinden en die een belangrijke rol spelen bij gezondheid en ziekte. Er worden steeds meer verbanden gevonden tussen verschillende microbenpopulaties en allerlei ziekten, waaronder obesitas en hart- en vaatziekten. Via *next-generation sequencing* kan men het geheel van DNA en RNA van de microben – het microbioom – in kaart brengen en zo bijvoorbeeld bepalen welke bacteriesoorten er op een bepaalde plaats aanwezig zijn.

Naast het microbioom zijn er nog veel meer omgevingsinvloeden. Deze zijn zeer divers van aard, van chemische stoffen in de lucht en voeding, fysische invloeden zoals straling, tot psychologische en sociale invloeden. Het leggen van relaties tussen gezondheid/ziekte en deze diversiteit van omgevingsfactoren, is het terrein van de epidemiologie. Ook op dit terrein speelt Big Data een steeds belangrijker rol. Gelet op de grote diversiteit aan omgevingsinvloeden is het onmogelijk deze uitputtend op te noemen. Daarom wordt dit onderwerp aan de hand van een aantal voorbeelden nader toegelicht.

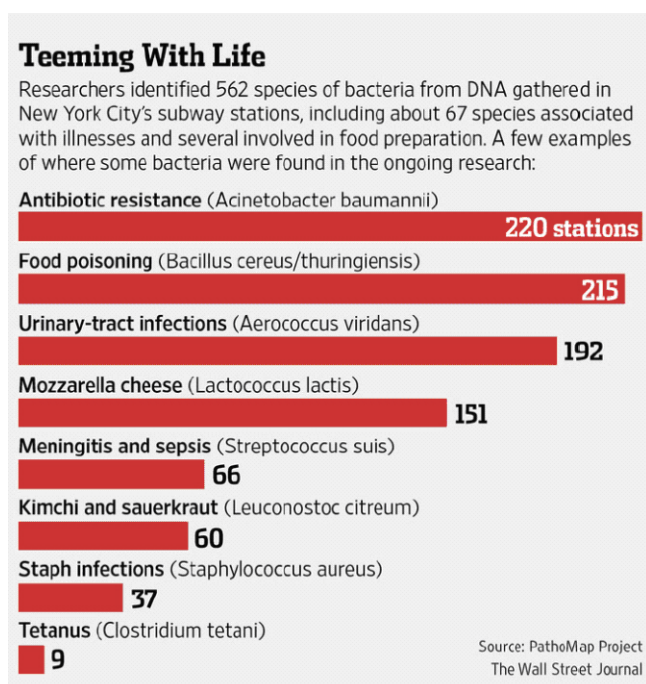
7.2 RELATIE ZIEKTE EN OMGEVINGFACTOREN

Onderzoek naar de relatie tussen ziekte en omgevingsfactoren kent een lange historie. Het ontdekken van verbanden is detectivewerk: systematisch analyseren van gegevens en het bij elkaar brengen van allerlei puzzelstukjes. Een klassiek voorbeeld is de ontdekking van Snow dat de cholera-epidemie in 1845 in Londen werd veroorzaakt door besmet water uit een waterpomp op Broadstreet door het verspreidingspatroon van de ziekte in kaart te brengen. De infectietheorie bestond in die tijd nog niet. De algemene theorie was dat cholera veroorzaakt werd door slechte lucht. Door zorgvuldige statistische analyse van de locatie van cholera-gevallen kon hij het verband tussen de kwaliteit van het bronwater en cholera aantonen. Bij epidemiologische studies worden de principes van Snow nog steeds toegepast.⁷⁹ De technieken die hierbij gehanteerd worden zijn wel verbeterd en ook hierbij wordt Big Data-analyse ingezet, zie kadertekst.

Bacteriën opsporen in de New Yorkse metro

Onderzoekers van het Weill Cornell Medical College in New York verzamelen in een lopend onderzoek bacteriesamples van bijvoorbeeld handvatten en deurknoppen, tourniquets en kaartjesautomaten in de 466 metrostations en treinstellen in verschillende delen van New York. Het DNA in de *samples* wordt gesequenced. De 10 miljard fragmenten worden door een supercomputer gesorteerd en vergeleken met een database van bekende bacteriën. De resultaten van het *PathoMap*-project laten veel 'problematische' bacteriën zien. Zo geeft onderstaande grafiek aan dat in bijna de helft van de stations antibioticaresistente bacteriën en bacteriën die voedselvergiftiging kunnen veroorzaken, voorkomen. Minder frequent komen veroorzakers van hersenvliesontsteking voor. Zelfs de bacterie die builenpest veroorzaakt is gevonden. Het bleek ook dat er verschillende bacteriepopulaties in verschillende wijken van New York voorkomen. Een belangrijke vraag is welke relatie dit heeft met andere ziektepatronen in verschillende wijken.

Figuur 7.1 Bacteriën in de New Yorkse metro



Bron: The Wall Street Journal, 2015⁸⁰

7.3 BEVOLKINGSONDERZOEKEN EN BIOBANKEN

De voorbeelden uit paragraaf 7.2 hebben betrekking op infectieziekten, maar zoals in de inleiding is aangegeven speelt de omgeving een belangrijke rol bij vele ziekten. Het is vaak lastig om vermeende verbanden hard te maken. Zo bestond in de jaren dertig van de vorige eeuw reeds het vermoeden dat asbest kankerverwek-

kend was. Dit was de reden dat in nazi-Duitsland het gebruik van asbest verboden was. Het duurde evenwel ruim zestig jaar voordat het gebruik van asbest in Nederland werd verboden (1993). Brede bevolkingsonderzoeken, waarbij gebruik gemaakt wordt van Big Data maken het mogelijk om potentieel ziekteverwekkende omgevingsfactoren sneller op te sporen. Een voorbeeld van een dergelijk onderzoek is Lifelines, een langlopend bevolkingsonderzoek naar het ontstaan en de ontwikkeling van chronische aandoeningen, zie kadertekst.

LifeLines

In de periode 2006 - 2013 zijn bewoners in het noorden van Nederland uitgenodigd deel te nemen aan een bevolkingsonderzoek. Inmiddels nemen 167.729 kinderen en volwassen hieraan deel. Dit cohort gaat ten minste 30 jaar lang gevolgd worden. Het bedrijf dat dit doet is LifeLines. Het Universitair Medisch Centrum Groningen is 100% aandeelhouder.

Elke vijf jaar krijgen de deelnemers een uitnodiging voor een uitgebreid medisch onderzoek. Zo worden onder andere bloeddruk, longfunctie en cognitief functioneren gemeten. Om de anderhalf jaar wordt de deelnemers gevraagd een enquête in te vullen. Men probeert de gegevensverzameling aan te vullen met gegevens uit medische dossiers en uit de omgeving, zoals luchtvervuiling.

Een belangrijk onderdeel van het LifeLines-project is de biobank, waarin bloed- en urinemonsters worden opgeslagen. Van elke deelnemer worden in 54 cupjes biomateriaal opgeslagen, zoals bloed, bloedplasma, serum en urine. Deze biomaterialen worden bij -80 graden Celsius in een biobank opgeslagen. Onderzoekers kunnen deze gebruiken voor wetenschappelijk onderzoek. Zo kan bijvoorbeeld het genoom bepaald worden en kunnen verbanden met in de tijd opgetreden ziekten onderzocht worden.

Doel van de gegevensverzameling is om de complexe wisselwerking tussen onder andere omgeving en genen te bestuderen en inzicht te krijgen in risicofactoren die van invloed zijn op ontstaan en beloop van chronische aandoeningen en hoe deze risico's per individu kunnen verschillen.

Bron: *lifelines.nl*

De LifeLines-biobank is een voorbeeld van een populatie-biobank. Er worden monsters van een representatieve doorsnede van de bevolking in opgeslagen. Het merendeel van de deelnemers is (vooralsnog) gezond.

Een ander 'soort' biobank is de klinische biobank. Hierin worden weefsels opgeslagen die zijn afgestaan door mensen die aan een bepaalde ziekte of aandoening lijden. Een voorbeeld hiervan is het Parelinoer Instituut, zie kadertekst.

Parelsnoer

Het Parelsnoer Instituut van de NFU omvat de biobanken van de acht Nederlandse universitaire centra. Deze centra verzamelen van dertien ziektes, de zogenoemde 'parels', biomaterialen en patiëntgegevens. De biomaterialen worden in de biobanken van de universitaire centra opgeslagen. Elk heeft één of meer parels onder zijn hoede. Zo is het Radboud umc projectleider van de 'parel' erfelijke darmkanker, het vumc van de ziekte van Parkinson en het ErasmusMC van ischemische hartziekten.

De biomaterialen zijn gekoppeld aan patiëntgegevens en worden voor wetenschappelijk onderzoek ter beschikking gesteld. Aanvragen worden door specialisten getoetst op zinvolheid en ook patiënten hebben een belangrijke stem.

Bron: Nederlandse Federatie van Universitaire Centra (NFU)

Ook in het buitenland zijn belangrijke initiatieven. Het *Million Veteran Program* in de VS is een voorbeeld hiervan, zie kadertekst.

Het Million Veteran Program

De Veterans Administration, onderdeel van het Amerikaanse ministerie voor veteranenzaken levert via meer dan 1.700 ziekenhuizen en andere zorginstellingen zorg aan ruim 8 miljoen veteranen.⁸¹ In 2013 is zij gestart met het *Million Veteran Project*. Genetische, *lifestyle*, werk- (met name verrichte militaire activiteiten) en zorginformatie van naar verwachting 1 miljoen veteranen zal in een immense database worden opgenomen. Van elke deelnemer wordt een buisje bloed afgenomen en opgeslagen voor toekomstige genetische analyse. Doel is om de invloed van genen en omgeving op ziekte te onderzoeken. Deelname is vrijwillig en staat geheel los van de zorgverlening en zorgverzekering. Daartoe zijn stringente beveiligings- en privacymaatregelen geïmplementeerd.

Bron: Veterans Affairs Office of Research and development⁸²

Biobanken en privacy

Biobanken met gerelateerde patiëntgegevens genereren enorme hoeveelheden gegevens met een grote variëteit. Deze Big Data is zeer waardevol voor wetenschappelijk onderzoek. De informatie is evenwel ook zeer privacygevoelig. Burgers/patiënten die lichaamsmateriaal voor onderzoek beschikbaar stellen moeten erop kunnen vertrouwen dat hun privacy gewaarborgd is en dat het materiaal gebruikt wordt waarvoor het bedoeld is. Zij geven, na vooraf geïnformeerd te zijn, hun toestemming, zogenoemd *informed consent*. Maar waarvoor zij precies toestemming geven is vooraf niet exact aan te geven. In de kadertekst over het Parelsnoerinitiatief is aangegeven dat aanvragen voor biomaterialen worden getoetst op zinvolheid door specialisten en dat ook patiënten een belangrijke stem hebben. Maar de mogelijkheid bestaat dat een burger/patiënt wier materiaal onderzocht gaat worden zich niet kan vinden in de beslissing van de toetsingscommissie. Via een *opt-out*-mogelijkheid zou de betrokkene zich moeten kunnen terugtrekken. Dit betekent echter dat deze voor elke onderzoek met zijn of haar lichaamsmateri-

aal apart toestemming moet geven, hetgeen een grote administratieve belasting van zowel de onderzoekers als patiënten is. En hoe moet met materiaal van overleden patiënten worden omgegaan? Zij kunnen geen specifieke toestemming meer geven. Een ander belangrijk punt is dat uit het onderzoek van het lichaamsmateriaal informatie naar voren kan komen die voor de betrokken patiënt relevant kan zijn, bijvoorbeeld over het al dan niet hebben van een al dan niet behandelbare erfelijke aandoening. Wil de patiënt hierover geïnformeerd worden?

7.4 PREVALENTIE EN INCIDENTIE VAN (INFECTIE)ZIEKTEN

Het bekendste voorbeeld van het gebruik van Big Data voor epidemiologische doeleinden is Google Flu c.q. Griep Trends, die de verspreiding van griep *realtime* volgde. Op de website werden de gegevens grafisch weergegeven (zie figuur 7.2). Op 20 augustus 2015 is Google hiermee gestopt.

Figuur 7.2 Griepactiviteit in de Verenigde Staten



Bron: Google Griep trends⁸³

Het model is gebaseerd op griepgerelateerde zoekvragen die gebruikers van Google invoeren. Ook voor Nederland houdt Google dit bij.⁸⁴ De klassieke manier voor het monitoren van de verspreiding van griep is met behulp van peilstations. Zo houden in Nederland sinds 1970 circa veertig huisartsenpraktijken de jaarlijkse griep epidemie bij.⁸⁵ In de Verenigde Staten houden de Centres for Disease Control and Prevention de gegevens bij. Deze officiële gegevens lopen gemiddeld twee weken achter op de realiteit. Zoals hiervoor aangegeven publiceert Google de gegevens niet meer. Zij verzamelt ze nog wel, maar levert ze nu door aan onderzoeksinstellingen, zoals aan de Columbia University's Mailman School, Boston Children's Hospital en de Centres for Disease Control and Prevention.⁸⁶

Een ander voorbeeld van het gebruik van Big Data is bij de verspreiding van cholera na de aardbeving in Haïti in 2010, zie kadertekst.

Twitter en de verspreiding van cholera

In 2010 was er een uitbraak van cholera in Haïti na de aardbeving. Twitter updates bleken even betrouwbaar als officiële rapporten over de verspreiding van de ziekte, maar waren wel twee weken eerder beschikbaar.

Bron: TechAmerica Foundation⁸⁷

Betrouwbaarheid van Big Data-analyses

In 2009 publiceerden onderzoekers van Google een artikel in *Nature*, waarin zij aangaven dat de resultaten van Google Flu Trends (GFT) voor 97% accuraat waren.⁸⁸ Echter Butler vond bij een follow-upstudie gepubliceerd in februari 2013 in *Nature* dat Google er in dat jaar flink naast zat en bijna twee keer zoveel griepgevallen meldde dan de officiële Centers for Disease Control and Prevention.⁸⁹ In een analyse geven Lazer et al in een artikel in *Science* aan dat de problematiek niet beperkt is tot GFT en wijzen op twee zaken die een rol speelden bij de (sterk) afwijkende resultaten, namelijk Big Data-overmoed en algoritme-dynamica.

Met Big Data-overmoed wordt bedoeld op de, veelal impliciete, aanname dat Big Data een alternatief is voor, in plaats van een aanvulling op, de traditionele wijze van gegevens verzamelen en analyseren. Big Data is vaak afkomstig van instrumenten die niet ontworpen zijn om valide en betrouwbare gegevens te genereren voor wetenschappelijke analyse. Bij de eerste versie van GFT hadden de ontwerpers getracht om een best match te vinden tussen 50 miljoen zoektermen en 1152 datapunten.⁹⁰ Dit maakt de kans op het vinden van matches die niets met griep te maken hebben vrij hoog. Dat het fout kon gaan bleek in 2009 toen een influenza-A (H1N1 - varkensgriep) Pandemie buiten het seizoen geheel gemist werd. Als grote voordeel van het gebruik van Big Data voor de griepvoorspellingen wordt de *realtime* beschikbaarheid van de informatie genoemd. Zo lopen de officiële griepgegevens van de Centers for Disease Control (CDC) twee weken achter op GFT. Lazer et al. geven echter aan dat extrapolatie van zelfs drie weken oude CDC-data een betere voorspelling geven dan de GFT-gegevens.⁹¹

Algoritme-dynamica betreft het in de loop van de tijd wijzigen van de gebruikte algoritmen. Zo paste Google het algoritme in 2009 aan, maar nadat bleek dat dit nieuwe algoritme het aantal griepgevallen systematisch overschatte en opnieuw aangepast moest worden. Google probeerde haar (geheime) zoekalgoritmen steeds verder te verbeteren, niet alleen als service naar de gebruikers, maar ook ter verbetering van haar *business model*, namelijk het verkrijgen van inkomsten uit advertenties. Daardoor worden zoekterm aantallen niet enkel beïnvloed door externe gebeurtenissen, zoals de griep. Dit betekent dat het lastig is gegevens stabiel en vergelijkbaar te verzamelen over een langere tijdsperiode.⁹²

Big Data en modetrends

Het griepvoorbeeld geeft aan dat men niet blind kan varen op Big Data, in dit geval zoekvragen. Dit geldt ook voor sociale media, omdat deze sterk aan ‘modetrends’ onderhevig zijn. Zo zijn volgens marktonderzoekers Facebook en Twitter sterk op hun retour bij de jongere generatie, terwijl de populariteit van Instagram en Pinterest zijn toegenomen.⁹³ Om trends te kunnen volgen zijn betrouwbare gegevens over een langere periode nodig. De vluchtigheid van sociale media maakt dit lastig. Daarnaast worden de gegevens vaak niet systematisch verzameld, zodat er gemakkelijk grote systematische vertekeningen, *bias*, op kunnen treden. Het grote probleem daarbij is dat deze moeilijk te detecteren en te corrigeren zijn.⁹⁴

7.5 OPSPOREN VAN BIJWERKINGEN VAN GENEESMIDDELEN

Voordat een geneesmiddel op de markt mag worden gebracht moet het uitgebreid getest worden in zogenoemde *clinical trials*. Hierbij worden vier fasen onderscheiden:

fase I: Het middel wordt getest bij een kleine groep vrijwilligers om te zien welke dosis veilig zijn en wat de bijwerkingen zijn.

fase II: Het middel wordt bij een grotere groep getest, waarbij gekeken wordt of het middel effectief en veilig is.

fase III: Het middel wordt bij een grote groep getest, waarbij naar de effectiviteit, ook in vergelijking met gangbare behandelingen, veiligheid en bijwerkingen wordt gekeken.

fase IV: Nadat het middel op de markt is gebracht en door grote groepen patiënten wordt gebruikt, worden de effecten op verschillende populaties en langetermijnbijwerkingen gemonitord (*post marketing surveillance*).

In fase IV, nadat het middel op de markt is gekomen en aan een brede patiëntenpopulatie wordt voorgeschreven, kunnen zeldzame bijwerkingen optreden die niet eerder zijn opgemerkt. Analyse van grote gegevensbestanden kan een belangrijke rol spelen bij de ontdekking hiervan zoals het onderstaande voorbeeld aangeeft, zie kadertekst.

Ontdekking van een gevaarlijke bijwerking van een geneesmiddel

In 1999 kwam de pijnstiller Vioxx op de markt. Het middel werd vooral voorgeschreven aan patiënten met reumatoïde artritis. Al snel ontstonden er verdenkingen dat het middel als bijwerking een hartinfarct of plotse hartdood kon geven.

Zorgverlener Kaiser Permanente analyseerde de gegevens van 1,4 verzekerden en kon de gevaarlijke bijwerkingen cijfermatig onderbouwen. Mede op basis van deze informatie werd het middel in 2004 van de markt gehaald.

Bron: FDA⁹⁵

7.6 VOEDSELVEILIGHEID

In steden als New York en Chicago maken inspecteurs gebruik van gegevens van *social media* om de kwaliteit van restaurants te bewaken, zie kadertekst.

Voedselinpecteurs controleren op geleide van Big Data

In Chicago controleren 32 voedselinpecteurs – *sanitarians* – de 15.000 restaurants in de stad. Zij kunnen niet alle restaurants bezoeken en selecteren dan ook de meest risicovolle. Dit gebeurt op klassieke wijze, op basis van risicofactoren. Deze factoren zijn opgesteld na analyse van gedurende zo'n tien jaar verzamelde gegevens. Gegevens zoals eerdere overtredingen en de locatie van het restaurant bleken minder van belang dan bijvoorbeeld het voorkomen van weersituaties waarbij ingrediënten sneller bederven. Tot voor kort waren de controleurs voor het registreren van voedselvergiftigingen vooral aangewezen op meldingen van restaurants zelf. Gebleken is dat mensen op sociale media melden als zij misselijk zijn en moesten braken. Er is een computerprogramma ontwikkeld dat Twitterberichten scant op woorden die gelinkt kunnen zijn met restaurantbezoek en voedselvergiftiging, zoals braken, overgeven, enzovoort. Degenen die een dergelijk bericht versturen wordt gevraagd om informatie en het verzoek hun ervaringen te rapporteren aan de inspectie.⁹⁶

New York maakt gebruik van het sociale netwerk Yelp⁹⁷. Op deze site kunnen mensenbeoordelingen plaatsen van winkels, restaurants of producten. Het New York Department of Health and Mental Hygiene (DOHMH) analyseert de 'crowd-sourced reviews' over restaurants. Net als in Chicago scant een computerprogramma de berichten op woorden die gelinkt kunnen zijn met voedselvergiftiging. Deze werden nader onderzocht. Uit een analyse gedurende negen maanden, waarbij 294.000 reviews werden gescreend, bleek dat er drie uitbraken waren gevonden die niet via de officiële kanalen waren gemeld.⁹⁸

Volgens een bericht in de Washington Post maakt men in New York inmiddels ook gebruik van Twitterberichten. Berichten die een indicatie kunnen geven van een mogelijke voedselvergiftiging tracht men te koppelen aan mogelijke berichten van dezelfde twitteraar enkele dagen daarvoor waarin deze melding maakt van een restaurantbezoek. Clustering van meerdere meldingen kunnen de inspecteurs naar een specifiek restaurant leiden.⁹⁹

7.7 HET INTERNET OF THINGS ALS BRON VOOR (EPIDEMIOLOGISCHE) GEGEVENS

Elektronica en sensoren worden steeds kleiner en goedkoper en wordt steeds meer voorzien van draadloze communicatiemogelijkheden. Een standaard *smartphone* beschikt over verschillende sensoren, zoals een camera, microfoon en versnellings-sensor. In auto's worden steeds meer sensoren ingebouwd die draadloos positie, snelheid en de toestand van verschillende onderdelen, zoals banden en schokbrekers, kunnen meten en via het internet doorgeven. Bij een ongeval kan de auto geheel zelfstandig hulp inroepen. Er zijn thermostaten voor de centrale verwarming die gegevens via wifi uitwisselen. De gebruiker kan daarmee waar ook ter wereld met een *app* op zijn of haar *smartphone* via het internet de temperatuur in huis instellen.

Dit geheel van ‘dingen’ die via het internet gegevens kunnen uitwisselen, wordt het *Internet of Things* (IoT) genoemd. Het aantal van deze apparaten neemt sterk toe en daarmee ook de hoeveelheid gegevens die de sensoren genereren en die via het internet gecommuniceerd worden.

De sensoren worden door fabrikanten in de apparaten aangebracht met een bepaald doel. Zo kan met een versnellingsensor gemeten worden dat een auto sterk hobbelt, een teken dat de schokbrekers mogelijk aan vervanging toe zijn. Het GPS-systeem wordt in eerste instantie in een auto geïnstalleerd om bij een ongeluk de positie van de auto te kunnen bepalen of, met name bij vrachtwagens, om de route en het rijgedrag van de chauffeur te monitoren. De combinatie van versnellingsgegevens en GPS kan echter ook voor geheel andere doeleinden gebruikt worden. Als veel auto's op een bepaalde weg heftig hobbelen, dan is er zeer waarschijnlijk sprake van een slecht wegdek, iets waarin een wegbeheerder geïnteresseerd is. Vroeger moest deze de wegen inspecteren, maar met het IoT kan deze de toestand van het wegennet vanachter een computer monitoren. Dit is een voorbeeld van één van de in paragraaf 2.4 genoemde kenmerken van Big Data, namelijk dat het vaak vanuit het ene domein wordt gebruikt voor beslissingen in het andere domein. Gecombineerd met het volume en de diversiteit van de door het IoT gegenereerde gegevens kan met recht van Big Data gesproken worden.

De door IoT gegenereerde data kunnen ook voor het verzamelen van epidemiologische gegevens gebruikt worden. Een tweetal voorbeelden zijn in onderstaande kaderteksten geschetst.

Fijnstof meten met de smartphone

De luchtkwaliteit is van invloed op de (volks)gezondheid. Mensen met aandoeningen van de luchtwegen kunnen ernstige klachten krijgen als er smog is.

Het RIVM, de GGD Amsterdam en de DCMR Milieudienst Rijnmond hebben een *app* voor *smartphones* ontwikkeld waardoor patiënten over *realtime* luchtkwaliteitsgegevens kunnen beschikken. De *app* toont de meetwaarden van fijnstof, ozon en stikstofdioxide die door circa 75 meetstations in Nederland worden verzameld. De 75 meetstations geven echter slechts een grof beeld van de luchtkwaliteit.

Om een fijnmaziger beeld te krijgen hebben het RIVM, het KNMI, de Nederlandse Onderzoeksschool voor Astronomie (NOVA) en het Netherlands Institute for Space Research (SRON) het ISPEX project opgezet. ISPEX meet fijnstof op een geheel nieuwe manier. Het maakt gebruik van de camera van een *iphone* waar een opzetstukje op wordt geklikt, zie foto. Hiermee kan het publiek de concentratie van fijnstof meten en samen met de locatiegegevens via het internet aanleveren. In 2013 en 2014 heeft een groot aantal vrijwilligers fijnstof op deze wijze gemeten.

Bron: *ispex.ispex.nl*

Radioactiviteit meten met de smartphone

Het RIVM heeft een vast meetnet, het Nationaal Meetnet Radioactiviteit (NMR), voor het meten van radioactiviteit in de vorm van gammastraling. Zo kan bijvoorbeeld een incident met een kerncentrale in kaart gebracht worden. Dit meetnet is, net als het meetnet voor fijnstof, vrij grofmazig.

De camera van een *smartphone* kan naast zichtbaar licht ook gammastralen opvangen. Door de camera met zwarte tape af te plakken wordt het licht buitengehouden en kan gammastraling gemeten worden. De meting met een *smartphone* is niet zo nauwkeurig als een meting met professionele apparatuur, maar de metingen kunnen met het professionele meetnet gecontroleerd worden. Het voordeel van *smartphones* is, dat er metingen op veel plaatsen gedaan kunnen worden, waardoor een gedetailleerder totaalbeeld ontstaat.

Het RIVM is in Zeeland een onderzoek gestart om met vrijwilligers straling met de *smartphone* te meten. De *app* stuurt de meetresultaten naar het RIVM in Bilthoven, die de metingen binnen een uur op een website zet. Gekozen is voor een start in Zeeland, omdat daar de meeste meetpunten van het NMR aanwezig zijn in verband met de kerncentrales in Borssele en (net over de grens in België) in Doel en de Centrale Opslag voor Radioactief Afval (COVRA). Zo kunnen de metingen van de vrijwilligers goed vergeleken worden met de officiële metingen.

Bron: Zeeuwse Milieu Federatie (zmf.nl)

8 BIG DATA EN KWALITEIT EN DOELMATIGHEID VAN DE ZORG

8.1 INLEIDING

Niet alleen zijn er, zoals in hoofdstuk 3 is beschreven, grote lacunes in de kennis over de effectiviteit van behandelingen in het algemeen, maar ook als het gaat om de resultaten van zorgverlening. Het meten van resultaten van zorg is moeilijk. Overleving is weliswaar gemakkelijk te meten, maar is slechts één aspect. Kwaliteit van leven is minstens zo belangrijk, maar is heel lastig in maat en getal te vangen. In veel gevallen wordt niet direct naar de resultaten gekeken, maar naar het proces aan de hand van (*evidence-based*) richtlijnen en protocollen. De aanname is dat de resultaten optimaal zullen zijn als richtlijnen en/of protocollen correct gevolgd zijn. *Evidence-based* richtlijnen en protocollen zijn evenwel gebaseerd op onderzoeksresultaten van groepen. Vrijwel niemand is exact even lang als de gemiddelde Nederlander, en de individuele patiënt zal ook zelden exact overeenkomen met de gemiddelde patiënt waarop de richtlijn of het protocol is gebaseerd.

Voor de bepaling van de doelmatigheid van de zorgverlening moet men naast de uitkomst in termen van overleving en kwaliteit van leven ook de (financiële) kosten van de zorg weten. Doelmatigheid is immers de verhouding tussen (financiële) kosten en de eerder genoemde opbrengsten. In dit hoofdstuk wordt een aantal voorbeelden beschreven van de inzet van Big Data om de kwaliteit en doelmatigheid van de zorg beter te kunnen meten en daarmee te verbeteren.

8.2 KWALITEIT EN VEILIGHEID VAN ZORG

Voor de bewaking van de kwaliteit en veiligheid van de zorg is een belangrijke rol weggelegd voor de Inspectie voor de gezondheidszorg (IGZ). Het toezicht van de IGZ omvat een breed terrein, van ziekenhuizen tot verpleeghuizen en GGZ. De uitvoering daarvan is risicogestuurd. Hiervoor is een instrument ontwikkeld, een zgn. dashboard, om risico's in de zorg te detecteren. (zie kadertekst).

Big Data voor selectie van risicovolle zorgverleners

Voor het zgn. het risicogestuurd toezicht heeft de IGZ een dashboard ontwikkeld. Het is een instrument om inspecteurs te helpen om risico's in de zorg te detecteren. Het dashboard bestaat uit een computerprogramma waarmee op een overzichtelijke en compacte wijze informatie wordt getoond die gerelateerd is aan kwaliteit en veiligheid van zorg. Veel van deze informatie is afkomstig van registraties waaraan instellingen gegevens leveren, zoals Dutch Hospital Data¹⁰⁰. Daarnaast zijn er nog andere bronnen zoals jaarverslagen. Trends in relevante gegevens, zoals ziekenhuissterfte, worden getoond. Al deze informatie vormt input voor het prioriteren en inrichten van het toezicht.

In 2014 vond in samenwerking met de Radboud Universiteit een onderzoek plaats naar de mogelijkheid van het gebruik van sociale media. Immers op sociale media beschrijven patiënten regelmatig hun ervaringen met de zorg.¹⁰¹ Een belangrijke website in dit kader is zorgkaartnederland.nl van de patiëntenfederatie NPCF. Op deze site staan veel beoordelingen van zorgaanbieders door patiënten. De patiëntervaringen zijn inmiddels geïntegreerd in het dashboard voor risicogestuurd toezicht.

8.3 DOELMATIGHEID VAN ZORG

Met name in de Verenigde Staten wordt Big Data ingezet voor kostenreductie. In het tijdschrift Health Affairs is een aantal onderzoeksresultaten gepubliceerd.¹⁰² Zo onderzochten Bates et al het management van patiënten met hoge zorgkosten. Vijf procent van de patiënten zijn verantwoordelijk voor de helft van de zorgkosten in de Verenigde Staten. Het is belangrijk deze te identificeren c.q. te voorstellen wie dit zijn, zodat zorgverleners tijdig kunnen behandelen.

Project Artemis

Binnen het project Artemis, ontwikkeld door het Institute of Technology van de Universiteit van Ontario, Canada, wordt de grote, continue datastroom van het monitoren van pasgeborenen op de *intensive care* geanalyseerd. Ziekenhuisinfecties kunnen daardoor 24 uur voordat zich symptomen openbaren voorspeld worden. Daardoor kan een behandeling eerder gestart worden, waardoor de overlevingskansen van pasgeborenen toenemen.

Bron: University of Ontario (<http://hir.uoit.ca/cms/?q=node/24>)

Een ander voorbeeld is het probleem van heropnames. Deze vormen een zware belasting voor de patiënt en zijn zeer kostbaar. Volgens de onderzoekers is een derde daarvan vermijdbaar. Op basis van Big Data kunnen zorgorganisaties algoritmen ontwikkelen om te voorspellen bij wie het risico voor heropname groot is, zodat tijdig geïntervenieerd kan worden c.q. de behandeling wordt aangepast. Hetzelfde geldt voor complicaties, zoals nierinsufficiëntie, infecties en bijwerkingen van geneesmiddelen. Door gebruik te maken van Big Data kan de zorg beter aangepast worden aan de individuele patiënt (*personalized medicine*).

Voor twee ander voorbeelden zie de kaderteksten. Alhoewel deze in de literatuur als voorbeelden van Big Data worden genoemd, is hierover discussie mogelijk. Ze hebben weliswaar betrekking op de verwerking van grote hoeveelheden gegevens, maar zouden ook als voorbeelden van *business intelligence* – het verzamelen en bewerken van gegevens binnen de eigen organisatie – gezien kunnen worden.

Data mining voor optimalisatie van knieoperaties en bloedtransfusies

Orthopedisch chirurgen in het Brigham and Woman's Hospital in Boston analyseerden grote hoeveelheden patiëntgegevens en isoleerden de factoren die bepalend zijn voor het succes van knieoperaties. Door het systematisch standaardiseren van de operaties nam de succeskans toe en namen de kosten af. Op eenzelfde wijze wist de Universiteit van Michigan de noodzaak voor toediening van bloedtransfusies met 31% te verminderen. Behalve in een vermindering van risico's voor de patiënt, resulteerde het in een kostenbesparing van \$200.000 per maand.

Bron: Healthcare IT News, 12 juni 2012¹⁰³

8.4 CONCLUSIE

Het gebruik van Big Data maakt een beter en doelmatiger toezicht op de kwaliteit van de zorgverlening mogelijk. Zorginstellingen kunnen hun zorgprocessen ermee optimaliseren. De kans van slagen van bijvoorbeeld operaties kan daardoor toenemen en de kosten kunnen afnemen. De voorbeelden die in dit hoofdstuk zijn beschreven geven aan dat Big Data een bijdrage kan leveren aan een betere kwaliteit en doelmatigheid van de zorgverlening.

9 DE GEVAREN EN BEPERKINGEN VAN BIG DATA IN DE ZORG

9.1 INLEIDING

In de vorige hoofdstukken is een aantal mogelijkheden aan de orde gekomen waarbij Big Data de kwaliteit en/of de doelmatigheid van de zorg kan verbeteren. Het gebruik van Big Data kent evenwel ook gevaren en beperkingen. Het belangrijkste gevaar, dat in het voorgaande reeds enkele malen is genoemd, is de mogelijke aantasting van de privacy van het individu. Een beperking van Big Data is in het voorbeeld van Google Flu aangestipt. Op deze aspecten wordt in dit hoofdstuk nader ingegaan.

9.2 BIG DATA EN PRIVACY

Medische c.q. zorggegevens hebben vaak betrekking op intieme zaken en zijn derhalve zeer privacygevoelig. Zoals in de vorige hoofdstukken is aangegeven, zijn deze gegevens, zoals genoominformatie, van cruciaal belang voor de vooruitgang van de wetenschap en de ontwikkeling van effectievere behandelingen. Het instrument voor de bescherming van de privacy van het individu bij het gebruik van zorggegevens voor onderzoeksdoeleinden is het verwijderen van de identificerende gegevens c.q. het anonimiseren van de gegevens. Door het ingenieus combineren van gegevens kan men echter de persoon achter de gegevens vaak toch achterhalen, zoals de kadertekst laat zien. Dit is een voorbeeld uit het ‘pre Big Data’-tijdperk, waarbij er maar twee bestanden gecombineerd behoeften te worden. Nu, in het Big Data-tijdperk, zijn er tientallen, zo niet honderden, gegevenspuzzelstukjes van vrijwel iedereen beschikbaar op het internet, die geautomatiseerd in elkaar gepast kunnen worden en zo een complete puzzel opleveren.

In een Big Data-omgeving is het zeer lastig om gegevens echt te anonimiseren. Ohm is zelfs van mening dat anonimisering in het Big Data-tijdperk in het geheel niet meer mogelijk is. Door het combineren van gegevens uit diverse bronnen is de kans groot dat gegevens gedeanonimiseerd kunnen worden.¹⁰⁴ Zo bleken veertig procent van de anonieme deelnemers aan een DNA-onderzoek via geavanceerde analyses toch geïdentificeerd te kunnen worden.¹⁰⁵ Degenen die stellen: ‘Privacy is a thing of the past’ lijken het gelijk aan hun zijde te hebben.¹⁰⁶

De-indentificatie van medische gegevens

Een overheidsorganisatie in Massachusetts, de Group Insurance Commission (CIG), kocht zorgverzekering in voor ambtenaren. Medio jaren negentig van de vorige eeuw stelde de CIG de geanonimiseerde declaratiegegevens van de verzekerden beschikbaar voor onderzoekers. De gegevens waren geanonimiseerd door het verwijderen van naam, adres en sociale verzekeringsnummer. Geboortedatum en postcodegebied stonden er nog wel

in. De gouverneur van Massachusetts, William Weld, verzekerde dat de privacy van de betrokkenen gegarandeerd was.

Sweeney ging op zoek naar de medische gegevens van de gouverneur, die zich in de geanonimiseerde gegevens moesten bevinden. Zij wist dat hij in Cambridge, Massachusetts woonde, een stad van 54.000 inwoners met zeven postcodegebieden. Voor twintig dollar kocht zij de stemlijst van deze stad. Dit bestand bevatte de namen, adressen, postcode, geboortedata en geslacht van alle stemmers. Door deze twee bestanden te combineren, bleken er zes personen in Cambridge te zijn die dezelfde geboortedatum hadden als de gouverneur. Drie daarvan waren man en één daarvan had de postcode van de gouverneur. Triomfantelijk stuurde zij de gouverneur zijn medische gegevens, inclusief diagnoses en voorgeschreven medicatie.

Bron: Greely (2007)¹⁰⁷

In zijn advies Patiënteninformatie¹⁰⁸ heeft de RVZ ervoor gepleit dat beveiliging en privacy al bij het ontwerp van informatiesystemen wordt meegenomen: *privacy-by-design*. Gelet op de snelle ontwikkelingen op het gebied van Big Data, is het uitermate belangrijk hierop te anticiperen en uit te gaan van *worst-case scenario's*.

9.3 GEGEVENSBEVEILIGING

In de Verenigde Staten brengen medische identiteitsgegevens op de zwarte markt tien tot twintig keer zoveel geld op dan creditcardgegevens, namelijk rond de 10 dollar per stuk in plaats van 0,5 - 1 dollar. De reden dat medische identiteitsgegevens zo gewild zijn, is dat diefstal ervan minder snel opgemerkt wordt, zodat criminelen er, soms jaren, lang misbruik van kunnen maken. Zo kreeg een patiënt rekeningen voor chirurgische ingrepen die hij niet had ondergaan, alsmede voor een scooter en andere zaken, die opliepen tot tienduizenden dollars.¹⁰⁹ Het bleek dat medische dossiers van een ziekenhuis waren gestolen. In de Verenigde Staten zijn zorgverleners verplicht beveiligingslekken die betrekking hebben op meer dan 500 mensen openbaar te maken. Met een zekere regelmaat worden deze gemeld, zoals een inbraak in computersystemen van Community Health Systems Inc., een grote aanbieder van ziekenhuiszorg in rurale gebieden van de VS, waarbij naar verluid Chinese hackers gegevens van 4,5 miljoen patiënten stalen. In januari 2015 werden gegevens buitgemaakt van 11 miljoen verzekerden bij Prenera Blue Cross. Een maand later was zorgverzekeraar Anthem aan de beurt. Daar werden gegevens van 78,8 miljoen verzekerden en medewerkers gestolen.

Voor Nederland zijn geen goede gegevens over *hacks* in de zorg bekend. Tot voor kort bestond geen verplichting tot openbaarmaking. Op 4 juni 2015 zijn de Wet bescherming persoonsgegevens en enkele andere wetten aangepast en is er een meldplicht voor datalekken. Deze moeten gemeld worden aan het College ter bescherming van persoonsgegevens (CBP).

9.4 INFORMATIE IS MACHT

In de vorige hoofdstukken is een aantal voorbeelden gegeven waarin het publiek, in de regel om niet, gegevens aanlevert.

Bij de voorbeelden in paragraaf 6.7 betreffende fijnstof en radioactiviteit verzamelen vrijwilligers actief gegevens. Het publiek wordt ingeschakeld omdat het, zoals het RIVM aangeeft, te kostbaar is om zelf overal metingen te doen. De gegevens die vrijwilligers aanleveren worden in deze voorbeelden door een publieke organisatie bewerkt en gratis openbaar gemaakt. Bij andere voorbeelden werd gebruik gemaakt van gegevens die door private partijen, zoals Google, Twitter en Facebook, beheerd worden. De informatie is in deze situaties actief aangeleverd door de gebruikers van deze diensten. Met de opkomst van het IoT worden gegevens steeds meer zonder tussenkomst van de gebruiker aan, veelal private, partijen aangeleverd. Een voorbeeld zijn locatiegegevens die *smartphones* doorgeven aan telecomproviders. Om gratis *apps* te kunnen gebruiken moet de gebruiker veelal toestemming verlenen voor toegang tot allerlei gegevens, zoals identiteits-, locatie- en contactgegevens, terwijl de *app* zelf niets van doen heeft met deze gegevens. Een verontrustend aspect hierbij is, dat indien de gebruiker toegang geeft tot bijvoorbeeld zijn contactgegevens, contactgegevens van derden mogelijk worden doorgegeven die daarvoor zelf geen toestemming hebben gegeven.

De vele 'gratis' diensten en *apps* op het internet worden bekostigd uit de verkoop van gegevens die door gebruikers of hun apparaten kosteloos zijn aangeleverd. In het digitale tijdperk is het min of meer de norm geworden dat bedrijven allerlei gegevens over gebruikers verzamelen en deze verkopen. Een voorbeeld hiervan is het abonnement op een internetaansluiting dat AT&T begin 2015 in de VS aanbood. Het abonnement voor snel internet bedroeg 70 dollar per maand. Daarbij wordt het internetgedrag van de gebruikers geregistreerd en doorverkocht. Wie dit niet wil, moet 29 dollar per maand extra betalen om niet gevolgd te worden.¹¹⁰ Met de opkomst van consumer-eHealth worden ook op het terrein van gezondheid en zorg steeds meer gegevens 'geupload naar de cloud'. Het kan hierbij gaan van de hoeveelheid lichaamsbeweging op een dag – stappentellers enzovoort – tot bloeddrukmetingen, bloedglucosewaarden, ECG's en foto's van verdachte plekken op de huid ('melanoom-*app*'): gegevens die commercieel interessant zijn. De vraag is, wie de beschikkingmacht over de gegevens heeft c.q. mag hebben.

9.5 BEPERKINGEN VAN BIG DATA

In de voorgaande hoofdstukken is reeds een aantal beperkingen van Big Data genoemd. Marcus en Davis hebben ze op een rij gezet en komen tot acht problemen met Big Data.¹¹¹

1. *Met Big Data kunnen verbanden aangetoond worden, vooral subtiele verbanden die met kleinere gegevenssets gemist kunnen worden. Het is echter niet zeker dat deze verbanden ook betekenis hebben.*

Als voorbeeld noemen ze het sterke positieve verband in de periode 1998-2007 tussen het aantal nieuwe gevallen van autisme en de verkoop van organisch voedsel. Het is evenwel zeer de vraag of het verband causaal is. Dit probleem geldt overigens niet alleen voor Big Data, maar voor alle correlaties die tussen gegevens worden gevonden. Het is een fundamenteel probleem in de epidemiologie. Echter naarmate meer (historische) gegevens c.q. Big Data geanalyseerd worden, zullen er meer correlaties gevonden worden, waarvan de betekenis niet bekend is.

2. *Big Data kan wetenschappelijk onderzoek niet vervangen.* Als voorbeeld noemen de auteurs het afleiden van de driedimensionale structuur van eiwitten op basis van de coderende DNA-sequentie. Data en statistische technieken alleen lossen dit niet op. Er zal altijd een analyse gebaseerd op fysische en chemische kennis nodig zijn. In het tijdschrift Wired stelde de hoofdredacteur Chris Anderson in een artikel getiteld 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete' precies het omgekeerde. Hierop kwam veel kritiek. Snijders, Matzat en Reips maakten duidelijk dat er bij Big Data vaak heel sterke aannames worden gedaan ten aanzien van de mathematische eigenschappen, maar dat deze in het geheel niet de realiteit van micro-processen behoeven weer te geven: "A crucial problem is that we do not know much about the underlying empirical micro-processes that lead to the emergence of the[se] typical network characteristics of Big Data".¹¹²

3. *De betrouwbaarheid van veel instrumenten die op Big Data gebaseerd zijn kan gemakkelijk ondermijnd worden.* Als voorbeeld noemen de auteurs het geautomatiseerd analyseren en beoordelen van essays van studenten. De computer let hierbij op twee factoren, die goed blijken te correleren met de beoordeling door menselijke beoordelaars, namelijk de lengte van de zinnen en het gebruik van 'moeilijke woorden'. Als studenten dit inde gaten hebben, zullen zij lange zinnen met veel obscure woorden gaan gebruiken.

Een concreet voorbeeld van manipulatie betreft Facebook. In juni 2014 verscheen een artikel in het tijdschrift *Proceedings of the National Academy of Sciences of the United States of America* met als titel *Experimental Evidence of Massive-Scale Emotional Contagion Through Social Networks*. Forbes onthulde

dat het een grootschalig psychologisch experiment betrof, dat Facebook had uitgevoerd op circa 689.000 gebruikers. Het bleek dat emoties doelgericht beïnvloed konden worden: “... Facebook has the ability to make you feel good or bad, just by tweaking what shows up in your news feed.”¹¹³

4. *Ook zonder manipulatie kunnen Big Data analyses onbetrouwbare gegevens opleveren, bijvoorbeeld doordat gegevens op verschillende manieren, met verschillende doelen verzameld zijn en/of niet consistent in de tijd zijn verzameld.* Een voorbeeld hiervan is Google Flu Trends, die in paragraaf 6.4 is besproken.
5. *Het ‘echo-effect’. Veel Big Data is afkomstig van het internet. Als de bron van informatie voor Big Data-analyse zelf het resultaat is van Big Data, kunnen er vicieuze cirkels ontstaan.* Als voorbeeld noemen de auteurs vertaalprogramma’s als Google Translate. Deze halen hun vertalingen uit identieke teksten in verschillende talen, bijvoorbeeld identieke Wikipedia-pagina’s in twee verschillende talen. Echter als deze Wikipedia-artikelen vertaald zijn met Google Translate, dan zal een initiële fout in Google Translate zijn weg vinden in Wikipedia en de fout in Wikipedia voedt weer Google Translate. Een ander voorbeeld is het peilen van het algemene consumentenvertrouwen. Het CBS heeft een proef gedaan, waarbij ze het sentiment in bijna een miljard sociale media-berichten vergeleken met het consumentenvertrouwen zoals dat op de klassieke manier wordt bepaald. Het sentiment via sociale media bleek goed te correleren met het consumentenvertrouwen. De cijfers van het consumentenvertrouwen hebben evenwel invloed op het sentiment via sociale media. Dus als dit op basis van sociale media wordt bepaald, dan ontstaat er een vicieuze cirkel.¹¹⁴
6. *Het gevaar van teveel correlaties.* Dit probleem sluit aan bij het onder punt 1 genoemde. Als er 100 keer gekeken wordt naar het verband tussen twee variabelen, dan zal er op basis van kansberekening gemiddeld vijf keer onterecht een statistisch significant verband gevonden worden bij een betrouwbaarheidsinterval van 95%. Dit sluit aan bij de in paragraaf 3.10 door Ioannidis geschetste problematiek bij wetenschappelijk onderzoek.
7. *Big Data lijkt wetenschappelijk onderbouwde antwoorden te geven op hopeloos vage vragen.* Als voorbeeld noemen de auteurs de twee pogingen om bekende personen te rangschikken in termen van zijn of haar historische importantie, de een het boek “Who is Bigger? Where Historical Figures Really Rank” van Steven Skiena en Charles Ward en het andere het MIT media lab project Patheon. Beiden gebruikten Wikipedia als bron en claimen wetenschappelijke precisie. Begrijpelijk(er)wijs komen beiden op een aantal dezelfde personen uit, zoals Jezus, Lincoln en Shakespeare, maar er zijn ook grote verschillen. Dat is ook niet verwonderlijk, want hoe concretiseer je ‘historische importantie’?

Een voorbeeld in de zorg is wellicht die met betrekking tot het meten van de kwaliteit van leven. Wat is kwaliteit van leven precies? Is het eenduidig te meten?

8. *Big Data werkt het beste bij zaken die frequent voorkomen, maar werkt niet in zeldzame situaties.*

Als voorbeeld worden geautomatiseerde tekstanalyses genoemd, die bijvoorbeeld gebruikt worden door zoekmachines of vertaalprogramma's. Hierbij wordt vaak gebruikgemaakt van trigrammen, drie woorden op een rij. Frequent voorkomende trigrammen kunnen statistisch gezien goed verwerkt worden, maar bij zeldzame lukt dat niet.

Voor de zorg betekent dit dat Big Data vooral van nut zal zijn bij frequent voorkomende aandoeningen, zoals hart- en vaatziekten en diabetes, maar minder effectief zal zijn bij bijvoorbeeld zeldzame erfelijke ziekten.

Marcus en Davis waarschuwen dan ook voor de hype dat Big Data worden gezien als oplossing voor alle problemen: "It's an important resource for anyone analysing data, not a silver bullet."

9.6 CONCLUSIE

Voorstanders van het gebruik van Big Data wekken vaak de indruk dat het een panacee voor alles is. Big Data biedt grote kansen, maar net zoals bij een geneesmiddel, heeft het ook ongewenste bijwerkingen. Zo is er het gevaar van inbreuk op de privacy en diefstal van gegevens. De privacy van burgers moet goed beschermd worden evenals de toegang tot gegevens. Hieraan gerelateerd is het punt van de beschikkingsmacht over gegevens. Informatie is macht. Veel zaken op het internet lijken gratis, maar gebruikers betalen, vaak zonder het te beseffen, met hun gegevens. Het voorbeeld van AT&T geeft aan dat het (extra) moeten betalen voor privacy, in de Verenigde Staten althans, reeds realiteit is.

Daarnaast kent Big Data zelf een aantal beperkingen, zoals het gegeven dat er correlaties mee aangetoond kunnen worden, maar niet of deze causaal of relevant zijn. Gegevens kunnen vrij gemakkelijk gemanipuleerd worden, en er kan *bias* en inconsistentie ontstaan. Big Data is een aanvulling op het bestaande en zeker geen vervanging.

10 BESCHOUWING EN TOEKOMSTPERSPECTIEF VAN BIG DATA IN DE ZORG

Een belangrijke vraag is hoe Big Data de zorg in de toekomst zal veranderen. Hoe zal de zorg er over twintig jaar uitzien? De afgelopen twintig jaar is de samenleving drastisch veranderd door de grootschalige toepassing van ICT: vrijwel iedereen heeft tegenwoordig toegang tot internet en maakt gebruik van diensten, zoals telebankieren, sociale media, enzovoort. De hoeveelheid aan en variëteit van de uitgewisselde gegevens zijn spectaculair toegenomen. Zij gaan nog sterker toenemen, omdat niet alleen mensen steeds meer gegevens genereren, maar vooral ook apparaten, van koelkasten, auto's tot zorgtoepassingen, die via sensoren allerlei zaken meten: het *Internet of Things*. Voor de zorg gaat het op dit moment, zoals in het voorgaande is beschreven, vooral om *lifestyle gadgets*, die onder andere de hoeveelheid lichaamsbeweging en hartslag meten. Op het terrein van sensoren zijn er grote ontwikkelingen: ze worden steeds kleiner, goedkoper en kunnen steeds meer zaken meten. Een andere grote bron van gegevens vormen, zoals beschreven, de *high-throughput*-technieken, die onder andere *next generation sequencing* mogelijk maken. Big Data heeft ook in de zorg zijn intrede gedaan. Grote bedrijven zoals Apple, Google, Samsung en IBM zien de zorg zelfs als een aantrekkelijke groeimarkt.

Big Data kan een grote bijdrage leveren aan hét grote probleem in de zorg: het grote tekort aan betrouwbare informatie. Alhoewel de kennis over ziekten en aandoeningen de afgelopen decennia sterk is toegenomen, is de precieze oorzaak, of het samenspel van oorzaken, van het merendeel van ziekten en aandoeningen niet bekend. Het grootste deel kan niet écht genezen worden, hetgeen resulteert in chronische aandoeningen, die vaak symptomatisch worden behandeld. Zoals in hoofdstuk 3 is aangegeven, is van de helft van de behandelingen het effect onbekend, omdat dit niet onderzocht is. En van de helft dat wel onderzocht is, is, zoals Ioannides heeft aangetoond, het resultaat veelal onbetrouwbaar. Integratie van grote hoeveelheden gegevens vanuit verschillende bronnen – genetische gegevens, het medisch dossier, biobanken, omgevingsfactoren, enzovoort – zal de kennis over de oorzaak van ziekten en aandoeningen sterk doen toenemen. Op basis van deze toegenomen kennis kunnen effectieve(re) behandelingen ontwikkeld worden.

Om dit toekomstperspectief succesvol te realiseren moet wel aan een aantal voorwaarden voldaan worden. Gegevens moeten gestandaardiseerd vastgelegd worden om ze te kunnen vergelijken. Daarnaast moeten gegevens beschikbaar zijn voor onderzoekers. Voor wat betreft gegevens die in het kader van de zorgverlening worden vastgelegd heeft de RVZ in zijn advies Patiënteninformatie uit 2014 gepleit

voor het Persoonlijk Gezondheidsdossier (PGD). Patiënten/burgers moeten via dit PGD kunnen beschikken over al hun gezondheidsgegevens. Daarnaast kunnen zij besluiten deze gegevens voor wetenschappelijk onderzoek beschikbaar te stellen. Medische gegevens zijn zeer privacygevoelig. In het hiervoor genoemde advies wordt dan ook gepleit om bij het ontwerpen van informatiesystemen rekening te houden met privacy: *privacy by design*. Regie van de overheid, niet alleen ten aanzien van privacy, maar ook wat standaardisatie betreft, is gewenst. Op dit moment wordt hier door middel van het zogenoemde informatieberaad invulling aan gegeven.¹¹⁵

In het vorige hoofdstuk is aangegeven dat het in het Big Data-tijdperk steeds lastiger wordt om gegevens echt te anonimiseren. Een belangrijke vraag is of privacy ‘*a thing of the past*’ gaat worden. Mensen lijken zich in de dagelijkse praktijk weinig zorgen te maken om hun privacy, gezien het gemak waarmee ze allerlei persoonlijke gegevens openbaar maken via sociale media. Daarnaast maakt vrijwel iedereen gebruik van ‘gratis’ diensten op het internet, zoals zoekmachines, e-mail, enzovoort. De gebruikers van deze diensten betalen echter met hun privacy, bijvoorbeeld in de vorm van informatie over de door hen geraadpleegde sites en worden bestookt met specifiek op hen gerichte reclame. Zoals gezegd maakt het gros van de mensen zich hierover geen zorgen. Het recht op privacy lijkt zich te verschuiven van iets waar men recht op heeft naar iets waarvoor men moet betalen. Hoe hoog deze prijs is, zal afhankelijk zijn van de marktomstandigheden: in een monopolistische markt kunnen hoge bedragen bedongen worden.

David Braue schrijft in zijn artikel ‘Gmail is boiling the frog – and we are the frog’ over het bedrijf Google:

“Google has positioned Gmail as a gateway drug to a world where everything runs according to Google. Google wants to manage our photos, our social media, our email, our word-processing documents, our everyday tasks, even our general documents.

It wants to harvest that information and feed it to the Borg that is far and away the largest single information-filtering service in the world. By continuing to buy into Google’s model, we are feeding the centralisation of human knowledge – one email at a time. This is the brave new world of the Internet, where privacy is an historical footnote and we are tricked or simply bribed to give it up.

By and large, we are quite happy to do so. We may not love the need to deliver our personal lives on a platter in exchange for a spam-free, easily-accessible and substantially awesome email experience – but we do so with a smile, over and over again.”¹¹⁶

Braue richt zich specifiek op het bedrijf Google, maar iets soortgelijks zou gezegd kunnen worden over andere bedrijven, zoals Facebook of Twitter.

Misschien is de teloorgang van de privacy nog niet eens het grootste probleem, maar wel wat Braue stelt 'de centralisatie van menselijke kennis'. In hoofdstuk 3 is het stellen van een medische diagnose door de computer besproken. De supercomputer Watson van IBM is daar genoemd, evenals de overnames door Google van een aantal *artificial intelligence*-bedrijven. Een vraag die in de toekomst prangend zal worden is: van wie is (medische) kennis? De breed gedragen opvatting binnen de wetenschappelijk wereld is dat kennis vrijelijk voor iedereen beschikbaar moet zijn. Internet heeft hieraan een belangrijke bedrage geleverd. Zo bieden topuniversiteiten gratis cursussen via het internet.¹¹⁷ Daarnaast worden steeds meer onderzoeksresultaten gepubliceerd in zogenoemde *open access journals*, zodat iedereen er kosteloos kennis van kan nemen.

Beschikbaarheid van gegevens is één, het nuttig kunnen gebruiken ervan om problemen in de praktijk op te lossen is twee. Daar zijn deskundigen voor, bijvoorbeeld de arts voor het stellen van een diagnose. De gegevens moet aan elkaar gerelateerd en geïnterpreteerd worden. *Artificial intelligence*-systemen zoals Watson gebruiken de (gratis) wetenschappelijke kennis en bouwen daarmee expertise op op een bepaald terrein, bijvoorbeeld medische diagnostiek. Dit kost geld en het ligt dan ook voor de hand dat voor het stellen van een diagnose door de computer een prijs wordt gevraagd, net zoals de expert dit doet. Net als bij experts bepaalt de mate van concurrentie de prijs. Bij experts zal er altijd concurrentie zijn, omdat één expert niet de gehele wereld zal kunnen bedienen. In het geval van een computer ligt dit anders. Net al bij de zoekmachines op het internet kan er een dominante partij komen, waardoor er een centralisatie van kennis – een alomvattende *corpus of knowledge* – ontstaat die in private handen is. De casus van Myriad Genetics in paragraaf 4.2 toonde de gevolgen van een privaat monopolie op een zeer beperkt terrein van de zorg. Het ging hier slechts om een tweetal genen, maar wat zullen de gevolgen zijn als het om brede medische kennisgebieden gaat?

Een probleem dat daarnaast kan ontstaan is dat niet alleen de menselijke kennis gecentraliseerd wordt en in private handen komt, maar dat ook de richting van kennisexploratie meer en meer privaat gestuurd gaat worden. Op dit moment wordt reeds veel klinisch onderzoek, met name *randomized controlled trials*, gefinancierd door de farmaceutische industrie. Dit zijn onderzoeken naar de effectiviteit van nieuwe geïmplementeerde geneesmiddelen. Begrijpelijkwijze financieren zij geen onderzoek naar niet-octrooieerbare geneesmiddelen of niet-farmacologische behandelingen. In hoofdstuk 6 is aangegeven dat grote bedrijven zoals Apple, Google, Samsung, IBM en Microsoft Big Data in de zorg als potentiële groeimarkt zien en bereid zijn hierin (fors) te investeren. In het voorgaande is aangegeven dat Big Data een grote bijdrage kan leveren aan het opvullen van de grote lacunes aan medische kennis. Een platform zoals de *ResearchKit* van Apple kan hierbij zeer nuttig zijn. De vraag is evenwel hoe zich dit alles in de toekomst zal gaan ont-

wikkelen en wat het gehanteerde verdienmodel zal zijn. Een veel gehanteerde strategie is om diensten eerst gratis aan te bieden en als er voldoende marktaandeel is opgebouwd en gebruikers niet meer zonder kunnen, er (veel) geld voor te vragen. Universiteiten zijn steeds meer afhankelijk geworden van derde geldstromen. Dit betekent dat de richting van (medisch) onderzoek steeds meer commercieel gestuurd zal gaan worden. Big Data zou dit proces kunnen versterken. Niet alleen de wetenschap, maar ook overheden kunnen hieraan ten prooi vallen. Zo heeft overheid allerlei taken in het kader van de monitoring van de volksgezondheid. Griep wordt gemonitord met behulp van peilstations, er is een meetnet voor het meten van fijnstof en radioactiviteit in de lucht, enzovoort. Dit is vrij kostbaar. In hoofdstuk 7 zijn verschillende voorbeelden gegeven hoe Big Data hierbij ingezet zou kunnen worden. Het is niet ondenkbaar dat de overheid in de toekomst uit financiële overwegingen steeds meer Big Data gaat inzetten die beheer(s)t wordt door private partijen. Hierdoor maakt de overheid c.q. de samenleving zich evenwel steeds meer afhankelijk van private partijen.

NOTEN

- 1 Humer, C, Finkle, J. Your Medical record is worth more to hackers than your credit card. Reuters 24 sept. 2014. www.reuters.com/article/2014/09/24/us-cybersecurity-hospitals-idUSKCN0HJ21120140924
- 2 Cerrato, P. Is Population Health Management the Latest health IT Fad? Information Week, 31 juli 2012
- 3 Cusack, H.G. The future state of clinical data capture and documentation: a report from AMIA's 2011 Policy Meeting. Journal of the American Medical Informatics Association p. 1-7, 2012
- 4 Hekster, big-data-deskundige bij IBM in: Big data in de zorg: niet alleen verzamelen. Medisch Contact 8 januari 2015, p.23
- 5 *The 'Omics Age*. Cambridge, MA: NPG Education, 2010. www.nature.com/scitable/ebooks/the-omics-age-16549951
- 6 Laney, D. 3D Management: Controlling Data Volume, Velocity, and Variety, Gardner Group, 2001
- 7 Soubra, D. The 3v's that Define Big Data, 5 juli 2012 www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data
- 8 www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data
- 9 Brumfiel, G. High-energy physics: Down the petabyte highway. Nature, January 19, 2011, 469, pp. 282-82.
- 10 www.ibmbigdatahub.com/infographic/four-vs-big-data
- 11 Transforming Health Care Trough Big Data, Strategies for leveraging big data in the health care industry. Institute for Health Technology Tranformation, 2013
- 12 Why are the 3 V's not sufficient to describe big data. <https://datafloq.com/read/3vs-sufficient-describe-big-data/166>
- 13 Understanding Big Data: the 7 V's. <http://dataconomy.com/seven-vs-big-data/>
- 14 Iafrate, F. From Big Data to Smart Data, ISTE Ltd, 2015 pagina 2
- 15 en.wikipedia.org/wiki/Moore%27s_law
- 16 Big data en privacy.Kamerbrief van de Minister van Economische Zaken, kenmerk DGETM-TM/14179608, 19 nov. 2014
- 17 Needham, J. Disruptive Possibilities. How Big Data Changes Everything. Scale Abilities Inc. O'Reilly Media Inc. 2013 pag. 4-5
- 18 Hadoop.apache.org
- 19 Iafrate, F. From Big Data to Smart Data, ISTE Ltd, 2015 pagina 5
- 20 En.wikipedia.org/wiki/Apache_Hadoop. Hadoop Hosting in the Cloud
- 21 Big Data in zicht. 10 oplossingen voor maatschappelijke impact met Big Data. Eindrapport 2014. Stichting de Nationale Denktank, www.nationale-denktank.nl/eindrapport2014
- 22 Transforming Health Care Trough Big Data, Strategies for leveraging big data in the health care industry. Pagina 6. Institute for Health Technology Tranformation, 2013
- 23 Nickerson, R.S. Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology, vol2 no2 p. 175-220, 1998

- 24 Tversky, A., Kahneman, D. Judgement under uncertainty: Heuristics and Biases. *Science* 185 (4157:124-1131), 1974
- 25 Wilson, T.D. et al. A new look at anchoring effects: Basis anchoring and its antecedent. *J. of Experimental Psychology: General* 125(4):387-402, 1996
- 26 Tversky, A., Kahneman, D. Judgment under Uncertainty: Heuristics and Biases. *Science*, New Series, Vol. 185, No. 4157. (Sep. 27, 1974), pp. 1124-1131.
- 27 Graber M, Gordon R, Franklin N. Reducing diagnostic errors in medicine: what's the goal? *Acad Med.* 2002;77:981-92.
- 28 Nendaz M, Perrier A. Diagnostic errors and flaws in clinical reasoning: Mechanisms and prevention in practice. *Swiss Med Wkly* 2012;142:w13706.
- 29 Berner, E.S., Graber, M.L. Overconfidence as a cause of diagnostic error in medicine. *Am.J.Med.* 121(5) S5-S23 (2008)
- 30 Croskerry, P. The importance of cognitive errors in diagnosis and strategies to minimize them. *Acad. Med.* 78(8), p.777-780 (2003)
- 31 www.nationaalkompas.nl/gezondheid-en-ziekte/ziekten-en-aandoeningen/zeldzame-ziekten/zeldzame-aandoeningen-samengevat/
- 32 www.dsm5.org
- 33 Os, van J. De DSM-5 voorbij! Persoonlijke diagnostiek in een nieuwe ggz, *Diagnosis Uitgevers*, Leusden, 2014
- 34 Shortliffe, E.H., Buchanan, B.G. Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. Addison-Wesley, Reading (1984)
- 35 Buchanan, B.G., Shortliffe, E.H. Rule-Based Expert Systems The Mycin Experiments of the Stanford Heuristic Programming Project, Addison Wesley 1984
- 36 en.wikipedia.org/wiki/Mycin
- 37 Su, Frances E. et al. "Medical Tests and Bayes' Theorem." *Math Fun Facts* www.math.hmc.edu/funfacts
- 38 Predictions made by Ray Kurzweil. *Wikipedia.org*
- 39 Gibbs, S. Google buys two more artificial intelligence startups. *The Guardian*, 23 oktober 2014.
- 40 Silver, D et al. Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484-489, 28 januari 2016
- 41 IBM Watson Sein erster Job ist Arzthelfer. *Chip. business.chip.de/news/IBM-Watson-Sein-erster-Job-ist-Arztthelfer_51623616*
- 42 Ferruci, D.A. Introduction to 'This is Watson', *IBM J. Res. Develop.* 56, 3-4 (2012)
- 43 Ioannidis, J.P.A. Why Most Published Research Findings Are False. *PLoS Med* 2(8):e124
- 44 Cochrane Collaboration
- 45 Lewis, S., Clarke, M. Forest Plots: trying to see the wood and the trees. *BMJ* 2001; 322:1479-1480
- 46 www.cochrane.org
- 47 BMJ Clinical Evidence. What conclusions has Clinical Evidence drawn about what works, what doesn't based on randomized controlled trial evidence? <http://clinicalevidence.bmj.com/x/set/static/cms/efficacy-categorisations.html>
- 48 <http://clinicalevidence.bmj.com/x/set/static/cms/efficacy-categorisations.html>

- 49 Advies Patiënteninformatie, Raad voor de Volksgezondheid en Zorg, 2014
50 Zie <http://genomemedicine.biomedcentral.com>.
- 51 Sanger, F. Nicklen, S., Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad Sci* 1977;74:5463-5467
- 52 Behjati, S., Tarpey, P.S. What is next generation sequencing? *Arch Dis Child Educ Pract Ed* 2013;98(6):236-238
- 53 New York Times 14 mei 2013
- 54 New York Times, 24 maart 2015
- 55 [En.wikipedia.org/wiki/Myriad_Genetics](http://en.wikipedia.org/wiki/Myriad_Genetics)
- 56 Regelado, A. EmTech: Illumina Says 228.000 Human Genomes Will Be Sequenced This Year. MIT Technology Review 2014.
- 57 Philips, K et al. Geneomic Sequencing: Assessing the health care system, policy and big-data implications, *Health Affairs* 33. no.7 (2014):1246-1253
- 58 Bolhuis, P. Signalement Preconceptioneel erfelijkheidsonderzoek. Forum Biotechnologie en Genetica 2014. www.forumbg.nl
- 59 Ropers, H. On the future of genetic risk assessment. *J. Community Genetics* 2012; 3:229-236
- 60 Bolhuis, P. Signalement Preconceptioneel erfelijkheidsonderzoek. Forum Biotechnologie en Genetica 2014. www.forumbg.nl
- 61 Mathijssen et al. Targeted carrier screening for four recessive disorders: High detection rate within a founder population. *Eur. J. of Med. Genetics* 58(2015)123-128.
- 62 Kingsmore, S. Comprehensive carrier screening and molecular diagnostic testing for recessive childhood diseases. *PLOS Curr.* 2012 May 2.
- 63 Advies Next generation sequencing in diagnostiek, Gezondheidsraad, 4 februari 2015
- 64 ZonMw - Systeembioogie en Systems Medicine www.zonmw.nl/nl/thema-detail/systeem-bioogie-systems-medicine/thema-detail
- 65 www.rivm.nl/Documenten_en_publicaties/Algemeen_Actueel/Nieuwsberichten/2013/Miljoen_65_plussers_loopt_risico_op_verkeerd_medicijngebruik
- 66 Hospital Admissions Related to Medication (HARM). Een prospectief, multicenter onderzoek naar geneesmiddel gerelateerde ziekenhuisopnames. Eindrapport. Division of Pharmacoeepidemiology & Pharmacotherapy, Utrecht Institute for Pharmaceutical Sciences, 2006.
- 67 Systeembioogie, NWO. www.nwo.nl/over-nwo/strategie/sperpunten+2011-2014/samenwerken+in+themas/themas+2007-2014/systeembioogie
- 68 Hood, L. Lezing tijdens het Metabolomics2010-congres. Amsterdam 2010
- 69 The Systems Medicine Web Hub is now online. ZonMW Nieuws, 10 april 2015.
- 70 Quantified Self Blog, oldest entries www.webcitation.org/66TEY49wv
- 71 In: Wolf, G. Quantified Self. www.webcitation.org/66TEHdz4d
- 72 Pappas, S. Expensive baby monitor give false reassurance, researcher says, *Livescience* 20 november 2014, www.livescience.com
- 73 Ottes, L. Consumenten-eHealth. A game-changer?!Achtergrondstudie bij het advies Consumenten-eHealth, Raad voor de Volksgezondheid en Zorg, 2015 www.rvz.net
- 74 Portable Ultrasound Machine www.mobisante.com
- 75 Wood, Alex. The internet of things is revolutionizing our lives, but standards are a must. *The Guardian*, 31 maart 2015

- 76 Apple presenteert ResearchKit. nieuwsbericht 10 maart 2015 Medisch Contact. medischcontact.artsennet.nl
- 77 Stanford launches smartphone app to study heart disease. med.stanford.edu, march 2015
- 78 Jacobs, F. Een Revolutie in medisch onderzoek - via je iPhone? smarthealth.nl, 19 maart 2015. www.smarthealth.nl
- 79 nl.wikipedia.org/wiki/John_Snow
- 80 Hotz, R.L. Big Data and Bacteria: Mapping the New York Subway's DNA The Wall Street Journal, 5 feb 2015
- 81 www.va.gov/health/
- 82 www.research.va.gov/mvp/veterans.cfm
- 83 www.google.org/flutrends/intl/nl/about/how.html
- 84 Griep trends. www.google.org/flutrends/nl/
- 85 www.nivel.nl/NZR/peilstations
- 86 Bowman, D. Google Flu Trends site shuts down. Signal data will go to Columbia, Boston Children's, CDC. FierceHealthIT, 24 augustus 2015. www.fiercehealthit.com/story/google-flu-trends-site-shuts-down/2015-08-24
- 87 Demystifying Big Data: A Practical Guide to Transforming the Business of Government. TechAmerica Foundation, Washington D.C. 2012
- 88 Ginsberg, J et al. Detecting Influenza epidemics using search engine query data. Nature 2009;457(7232):1012-4
- 89 Butler, D. When Google got flu wrong. Nature 2013;494(7436):155-6.
- 90 Ginsberg, J. et al. Detecting influenza epidemiology using search engine query data. Nature 457, 1012 (2009)
- 91 Laser et al. The Parable of Google Flu: Traps in Big Data Analysis. Science 2014;343:1203-5
- 92 Ginsberg, J. et al. Detecting influenza epidemiology using search engine query data. Nature 457, 1012 (2009)
- 93 Oosterveer, D. Social media in 2015: Jongeren haken af van Facebook en Twitter. Marketingfacts. www.marketingfacts.nl/berichten/social-media-in-nederland-2015-jongere-haken-af-van-facebook-en-twitter
- 94 Davis, E. Geciteerd in : For Big Data, Big Questions Remain. Health Affairs, juli 2014 p.1113
- 95 www.fda.gov/NewsEvents/Testimony/ucm113235.htm
- 96 Harris, J.K. Health Department Use of Social Media to Identify Foodborne Illness – Chicago, Illinois, 2013-2014. Centers for Disease Control and Prevention, Morbidity and Mortality Weekly Report August 15, 2014 / 63(32):681-685.
- 97 www.yelp.com en www.yelp.nl
- 98 Harrison, C et al. Using Online Reviews by Restaurant Patrons to Identify Unreported Cases of Foodborne Illness – New York City, 2012-2013. Centers for Disease Control and Prevention, May 23, 2014 / 63(20):441-445.
- 99 Ravindanath, M. In Chicago, food inspectors are guided by big data. The Washington Post, 28 september 2014.
- 100 www.dhd.nl

- 101 Reshape blog. [Radboudreshapecenter.com/blog/innovatie-van-ongekende-kracht-neder-](http://Radboudreshapecenter.com/blog/innovatie-van-ongekende-kracht-nederland)
land.
- 102 Health Affairs, 8 juli 2014
- 103 Healthcare IT News, 12 juni 2012
- 104 Ohm, P. Broken promises of privacy: responding to the surprising failure of anonymisation.
UCLA Law Rev. 2010;57;1701-77
- 105 Tanner, A. Harvard professor reidentifies anonymous volunteers in DNA study. Forbes
25 april 2013
- 106 Why privacy is a thing of the past. [www.theguardian.com/technology/2013/oct/06/big-](http://www.theguardian.com/technology/2013/oct/06/big-data-predictive-analytics-privacy)
data-predictive-analytics-privacy
- 107 Greely, H.T. The Uneasy Ethical and Legal Underpinnings of Large-Scale Biobanks. 8 Ann.
Reva. Genomics. Genetics 343, 352 (2007)
- 108 Advies Patiënteninformatie, Raad voor de Volksgezondheid en Zorg, 2014
- 109 Humer, C, Finkle, J. Your Medical record is worth more to hackers than your credit card.
Reuters 24 sept. 2014. [www.reuters.com/article/2014/09/24/us-cybersecurity-hospitals-](http://www.reuters.com/article/2014/09/24/us-cybersecurity-hospitals-idUSKCN0HJ21120140924)
idUSKCN0HJ21120140924
- 110 Dwoskin, E., Gryta, T. AT&T Offers Data Privacy - For a Price. Wall Street Journal, 18 feb.
2015.
- 111 Marcus, G., Davis. E. Eight (No, Nine!) Problems With Big Data, The New York Times
<http://nyti.ms/1kgErs2>
- 112 Snijders, C. Matzat, U., Reips, U.-D. Big Data: Big gaps of knowledge in the field of Internet.
International Journal of Internet Science, 2012;7:1-5 [www.ijis.net/ijis7_1/ijis7_1_edi-](http://www.ijis.net/ijis7_1/ijis7_1_editorial.html)
tiorial.html.
- 113 McNeal, G.S. Facebook Manipulated Users News Feeds to Create Emotional Responses,
28 juni 2014, Forbes.com
- 114 Sandijk, J. CBS onderzoekt de mogelijkheden van big data. CBS, zomer 2013, 03/13 p.30-33
- 115 Een duurzaam informatiestelsel voor de zorg. Kamerbrief van de minister van VWS, 23 okto-
ber 2013, kenmerk 154772-110935-MC.
- 116 Braue, D. Gmail is boiling the frog - and we are the frog. ZDNet, 15 november 2013
- 117 Zie bijvoorbeeld www.edx.org

