Prof.dr. S. le Cessie

Toeval en vertekening

Universiteit Leiden

Toeval en vertekening

Oratie uitgesproken door

Prof.dr. S. le Cessie

bij de aanvaarding van het ambt van hoogleraar in de

Statistische Methoden in Observationeel (klinisch) Epidemiologisch Onderzoek

aan de Universiteit Leiden

op vrijdag 22 september 2017

Universiteit
Leiden

Meneer de Rector Magnificus, geachte aanwezigen,

Geeft roken een hoger risico op dementie? Leiden drie koppen koffie per dag tot minder trombose? Moet een bevalling ingeleid worden bij een groeivertraging? Waarom krijgen sommige mensen met overgewicht last van diabetes, kanker en andere aandoeningen terwijl anderen gezond oud worden? Allemaal vragen over oorzaak en gevolg, over causaliteit. Medisch onderzoek probeert het antwoord op dit soort vragen te vinden. Vaak wordt dat gedaan door grote groepen mensen te bestuderen, wie rookt er, hoeveel koffie drinkt men, wie wordt er ziek, hoe verloopt de ziekte? Dit type onderzoek wordt epidemiologisch onderzoek genoemd. Epidemiologen voeren dit soort onderzoek uit en denken na over de beste manier om dit te doen.

Wat is mijn rol in dit soort onderzoek? Ik ben een statisticus en houd me vooral bezig met methoden om epidemiologisch onderzoek op te zetten en te analyseren. Om uit dit soort onderzoek de juiste conclusies te trekken is niet eenvoudig. Resultaten kunnen door toeval afwijken of vertekend zijn. Ik ga u hier meer over vertellen. Over statistische methoden waarmee oorzaak en gevolg gemodelleerd kunnen worden. Over nieuwe bronnen van gegevens waarmee epidemiologisch onderzoek kan worden uitgevoerd, *big data*. En natuurlijk zal ik ook de medische statistiek zelf bespreken, waarbij ik u wil schetsen hoe boeiend het is om als statisticus in een medische omgeving te werken.

Gerandomiseerd onderzoek

Hoe kunnen we nu bepalen of roken, koffiedrinken, of een behandeling daadwerkelijk een effect heeft op een bepaalde uitkomst? Soms wordt dit soort vragen beantwoord met gerandomiseerd onderzoek. Bij gerandomiseerd onderzoek wordt door loting bepaald welke behandeling een patiënt krijgt, natuurlijk met instemming van de patiënt. Door de loting zijn de patiënten in de verschillende behandelingsgroepen vergelijkbaar qua prognose. Verschillen in uitkomsten tussen de groepen kunnen berekend worden, waarbij de statistiek nodig is om de onzekerheid in de bevindingen weer te geven. Wanneer we verschillen zien die groter zijn dan door toeval verwacht, kunnen we die toeschrijven aan het verschil in behandeling.

Gerandomiseerd onderzoek wordt vaak gezien als de meest valide vorm van onderzoek. Maar de kosten van het uitvoeren zijn vaak hoog. De patiëntenaantallen zijn daarom ook vaak relatief klein, en resultaten zijn meestal pas jaren later bekend. In de praktijk wordt ook lang niet iedere geschikte patiënt geïncludeerd in een studie. Soms heeft de patiënt een sterke voorkeur voor een behandeling en soms de dokter. Dat kan leiden tot selectieve patiëntengroepen, die afwijken van de dagelijkse klinische praktijk. Een voorbeeld hiervan is de Digitat[1] studie die een aantal jaar geleden uitgevoerd is door professor Sicco Scherjon en doctor Kim Boers, waarbij ik als statisticus nauw betrokken was. In deze studie werden vrouwen die minimaal 36 weken zwanger waren met een verdenking op een groeiachterstand bij het kind, gerandomiseerd tussen inleiden van de bevalling of afwachten, met zorgvuldige controles. In de gerandomiseerde studie gaven beide behandelingen equivalente uitkomsten bij de kinderen. Maar lang niet elke vrouw die gevraagd werd deel te nemen aan deze studie stemde ook toe. Een sterk punt van de Digitat studie was dat ook van de vrouwen die niet gerandomiseerd wilden worden, op dezelfde manier gegevens verzameld werden. De overgrote meerderheid van deze vrouwen koos voor niet inleiden, waarschijnlijk hopend op een natuurlijke bevalling. Juist bij deze weigeraars, die gemiddeld iets ouder en hoger opgeleid waren en daarbij minder rookten en minder zwaar waren, kwamen meer slechte uitkomsten voor. Verschillende verklaringen zijn hiervoor denkbaar: toeval, een minder goede monitoring buiten studieverband, of het feit dat deze vrouwen anders zijn, de voorkeur om natuurlijk te bevallen kan er voor gezorgd hebben dat er langer gewacht werd met ingrijpen. Het langer afwachten kan leiden tot een verhoogd risico op sterfte en minder optimale neurologische ontwikkelingen. De waarnemingen buiten het

gerandomiseerde onderzoek gaven in deze studie dus extra relevante informatie, die meegenomen is in de uiteindelijke aanbevelingen.

Observationeel onderzoek

Er zijn ook veel vragen die niet met gerandomiseerd onderzoek te beantwoorden zijn. Onze vraag over het mogelijke verband tussen roken en dementie bijvoorbeeld. We kunnen moeilijk zestienjarigen aanmoedigen om te gaan roken om dan te kijken of ze zestig jaar later dement zijn geworden. Een grote uitdaging voor statistici en epidemiologen is om ook in situaties waar geen gerandomiseerde data beschikbaar is, met redelijke betrouwbaarheid uitspraken te doen over mogelijke causaliteit. We gebruiken dan gegevens uit observationeel onderzoek: onderzoek waarbij er geen interventies uitgevoerd worden, maar waar alleen de huidige praktijk geregistreerd wordt.

In de afdeling klinische epidemiologie, waar ik werkzaam ben, wordt veel observationeel onderzoek opgezet en uitgevoerd. Een voorbeeld is de NEO-studie[2], waar Renée de Mutsert de dagelijkse leiding over heeft. Deelnemers aan de NEO-studie zijn tussen de 45 en 65 jaar en woonachtig in Leiden en omgeving. Een groot aantal van hen heeft overgewicht. Deze mensen hebben vragenlijsten ingevuld over hun gezondheid, bewegings- en voedingspatroon. Er zijn een groot aantal metingen uitgevoerd waaronder bloedafnames, vetmetingen en soms een MRI-scan. Het doel van de NEO-studie is om te begrijpen welke processen bij mensen met overgewicht er toe leiden dat sommigen van hen obesitas-gerelateerde ziekten krijgen, terwijl anderen ondanks het overgewicht gezond blijven. Alle deelnemers worden nu ook gevolgd om de ontwikkeling van de gezondheid en het ontstaan van ziekten te registreren.

Wanneer we uitspraken willen doen over oorzaak en gevolg met observationele gegevens, moeten we oppassen voor vertekening. Wanneer we bijvoorbeeld observeren dat koffiedrinkers minder vaak trombose krijgen, hoeft dat niet door de koffie veroorzaakt te worden. Koffiedrinkers zijn anders dan thee- of waterdrinkers. Zo zijn er relatief meer mannen onder de koffiedrinkers, en mannen hebben een wat lager risico op trombose. Dit fenomeen dat groepen niet direct vergelijkbaar zijn omdat er andere factoren meespelen, factoren die zowel de blootstelling of behandeling als de uitkomst beïnvloeden noemen we confounding. Om die kluwen van factoren te ontrafelen, en het echte effect van koffiedrinken, of roken of een behandeling te bepalen, zijn allerlei statistische methoden ontwikkeld. Vaak gebruiken we regressiemodellen zoals het logistische model of het Cox model om voor confounding te corrigeren. Maar deze regressieaanpak is niet altijd waterdicht. Bij heterogeniteit van effecten bijvoorbeeld, of wanneer we behandelingen over de tijd heen bestuderen, waarbij confounders van waarde veranderen als gevolg van eerder gegeven behandelingen. We moeten dan gebruik maken van speciale causale methoden, ik bespreek die zo verder.

Vertekening door selectie is een ander probleem in observationeel onderzoek. Dit speelt bijvoorbeeld wanneer we een specifieke subgroep volgen zoals patiënten met een bepaalde ziekte, personen met overgewicht, ouderen of patiënten in het ziekenhuis. Dit fenomeen is voor het eerst beschreven door de statisticus Berkson.[3] Hij laat zien dat factoren die in de algemene bevolking niet samenhangen (als voorbeeld gebruikt hij diabetes en galblaasontsteking), in een ziekenhuispopulatie opeens wel gecorreleerd kunnen zijn. Dat komt door de selectie die is gemaakt: mensen zonder diabetes en galblaasontsteking worden minder vaak opgenomen in een ziekenhuis. Selectiebias kan zorgen voor paradoxale bevindingen. Zo bleek in een onderzoek dat kinderen met een laag geboortegewicht waarvan de moeder tijdens de zwangerschap heeft gerookt het gemiddeld beter te doen dan kinderen met een laag geboortegewicht van niet-rokende moeders.[4] Dit is zeker geen causaal verband, we moeten zwangere vrouwen niet gaan aansporen om te roken. Het is te verklaren doordat roken tijdens de zwangerschap juist het geboortegewicht omlaag brengt en we hier selecteren op laag geboortegewicht. Vertekening door selectie speelt op heel veel gebieden: in onderzoek naar recidief, we spreken

dan van index-event bias, in onderzoek naar patiënten met en zonder overgewicht: de obesitas paradox, en in onderzoek naar ouderen, survival bias. Samen met doctor Anna Boef en professor Olaf Dekkers heb ik laten zien dat dit fenomeen ook speelt in genetisch onderzoek bij ouderen, resultaten van mendeliaanse randomisatiestudies in ouderen kunnen vertekend zijn.[5] Promovendus Roelof Smit bestudeert nu hoe groot deze vertekening kan zijn in een aantal realistische situaties.

Causale modellen

Hoe kunnen we nu rekening houden met deze bronnen van vertekening in observationeel onderzoek? In de afgelopen twintig jaar is hier een keur aan statistische en epidemiologische methoden voor ontwikkeld. Ik noem de Directed acyclic graphs (DAG), diagrammen waarmee assumpties over oorzaak en gevolg grafisch weergegeven worden. Met deze causale diagrammen kan vertekening door confounding en selectiebias geïdentificeerd worden en ze geven aan voor welke variabelen er wel en niet gecorrigeerd moet worden in de statistische analyse.

Naast deze DAGs is het mogelijk om causale effecten te schatten uit observationele data door causaal te modelleren. Ik ga hier niet in op de wiskundige details, hoe mooi die ook zijn, maar zal de grote lijnen schetsen. Het uitgangspunt is een gedachtenexperiment: hoe zou je, onder ideale omstandigheden in een ideale onderzoekswereld waar we interventies kunnen uitvoeren zonder ethische beperkingen en zonder weigeraars en uitvallers, de perfecte studie uitvoeren om de onderzoeksvraag te beantwoorden? Interventies zijn in mijn hypothetische wereld breed gedefinieerd: behandeling, medicatie maar ook genen, roken en vetverdeling vallen in mijn wereld onder interventie. De eerste stap is heel nauwkeurig de interventie te definiëren. Zijn we bijvoorbeeld geïnteresseerd in het effect van het voorschrijven van een behandeling, het effect van het daadwerkelijk starten met de behandeling, of het effect van de behandeling als deze gedurende drie maanden volgens voorschrift gevolgd is? Vervolgens kiezen we de uitkomstmaat, waarmee we het effect van de interventie willen meten. Daarna moet ook de populatie waarin we het effect willen bestuderen (de hele bevolking, jongeren die op 16-jarige leeftijd met roken beginnen) nauwkeurig gedefinieerd worden.

In mijn perfecte studie zouden we iedere patiënt alle verschillende behandelingen willen geven. Misschien kent u de film Groundhog Day, waar een weerman in een tijdlus belandt, en elke ochtend opnieuw weer wakker wordt op 2 februari, in Punxsutawney, Pennsylvania, en keer op keer dezelfde dag beleeft? Dat wil ik in mijn ideale studie ook, de klok terugdraaien, elke patiënt onder dezelfde omstandigheden elk van de interventies geven en steeds opnieuw kijken wat de uitkomst is. Die mogelijke uitkomsten noemen we potentiele uitkomsten. In de praktijk zien we natuurlijk per patiënt slechts het effect van een van de interventies. De andere uitkomsten zijn *counterfactual*, uitkomsten die we zouden zien als er anders gehandeld was. Toch is het concept van potentiële uitkomsten heel bruikbaar want daarmee kunnen we precies het causale contrast waarin we geïnteresseerd zijn definiëren. Wanneer we bijvoorbeeld willen weten wat de gezondheidswinst is als pubers niet meer beginnen met roken, dan zijn we in causale termen geïnteresseerd in het *treatment effect in de treated*, het verwachte verschil in potentiele uitkomsten tussen wel en niet roken, in personen die als puber met roken gestart zijn. Wiskundig kunnen we dit verschil precies in formules opschrijven.

De volgende stap is om het gewenste causale contrast dat we wiskundig exact geformuleerd hebben te schatten met de beschikbare data. Hiervoor is een heel instrumentarium aan statistische methoden beschikbaar. Met deze methoden kunnen we, gebruik makend van observationele gegevens, veel verschillende relevante problemen aanpakken. We kunnen subgroepen identificeren die veel of juist weinig baat bij de interventie hebben, personalized medicine. We kunnen behandelingsstrategieën door de tijd heen vergelijken, bijvoorbeeld om te bepalen wat het beste moment is om een behandeling te starten. We kunnen causale methoden gebruiken om ziekte-

processen beter te begrijpen, zoals in de NEO-studie waar we met mediatieanalyse de verschillende paden van overgewicht naar ziekte bestuderen.

Kortom, deze methoden zijn in een groot aantal verschillende toepassingsgebieden erg bruikbaar. Dat laat onverlet dat de naam *causale modellen* niet garandeert dat deze methoden valide conclusies opleveren. Bij het schatten van de effecten worden er altijd aannames gemaakt. Zo veronderstellen veel causale schattingsmethoden dat alle confounding gemeten is, iets wat in de praktijk vaak te betwijfelen valt. Er moet variatie in blootstelling of behandeling zijn**.** Wanneer bepaalde patiëntengroepen strak volgens richtlijnen behandeld worden, zijn de methoden niet bruikbaar. Meestal zijn er grote databestanden nodig. Hoe specifieker we interventies definiëren, hoe meer data nodig zijn. Als we het effect van twintig jaar roken willen bestuderen, moeten er gegevens zijn van mensen (met verschillende confounderwaarden) die daadwerkelijk zo lang gerookt hebben.

Bij een goede causale analyse hoort in ieder geval een methodologische bijsluiter waarin alle assumpties staan die er tijdens de analyses gemaakt zijn. Het is belangrijk om te kijken hoe robuust de resultaten zijn en er moeten sensitiviteitsanalyses uitgevoerd worden om de effecten van bijvoorbeeld meetfouten en ongemeten confounding te bepalen. Voor mediatieanalyse hebben wij daar al methoden voor ontwikkeld.[6,7]

Er liggen nog veel uitdagingen in dit causaliteitsonderzoek, zowel in de methodologie als in het toepassen ervan. Omgaan met veel confounders bijvoorbeeld, met confounders met meetfouten, met missende observaties. Het ontwikkelen van goede sensitiviteitsanalyses en betere schattingsmethoden. En onderzoek naar alternatieve causale methoden wanneer er ongemeten confounding is, bijvoorbeeld door gebruik te maken van verschillen in behandelingsvoorkeuren tussen behandelaars of ziekenhuizen (instrumentele variabelen) of van verandering in behandeling over de tijd.

Nu even terug naar het gerandomiseerde onderzoek. Impliceren de ontwikkelingen in het causaal modelleren nu dat we geen gerandomiseerd onderzoek meer hoeven uit te voeren? Dat zeker niet! Het grote voordeel van randomisatie is dat niet alleen gemeten maar ook alle ongemeten factoren die de uitkomst beïnvloeden gelijk verdeeld zijn over de verschillende behandelingsarmen. Maar er valt veel winst te behalen door gerandomiseerde trials in te bedden in observationele cohorten en registraties, zoals in de Digitat-studie gedaan is. Observationele data kunnen input genereren voor de trials, wat tot kleinere steekproefgroottes kan leiden. De resultaten zijn ook beter generaliseerbaar naar de dagelijkse praktijk, en vragen over variatie in dosering tussen subgroepen kunnen beantwoord worden. En ook hier liggen nog veel relevante methodologische vragen open.

Hergebruik van routinematig verkregen data

In veel medische studies zoals de NEO-studie en de Digitat-studie, worden gegevens speciaal voor onderzoek verzameld. Vooraf wordt nagedacht over de onderzoeksvragen en het studiedesign. In het studieprotocol wordt precies vastgelegd welke gegevens er verzameld gaan worden en op welke tijdsmomenten dat gebeurt. Elke proefpersoon wordt op dezelfde manier, objectief, nauwkeurig en volledig gemeten en op dezelfde manier gevolgd in de tijd.

Een alternatief voor deze werkwijze is om gegevens die voor andere doeleinden verzameld zijn te gebruiken voor onderzoek. Zo genereert de patiëntenzorg grote hoeveelheden data. In elke spreekkamer staat een computer waarmee de arts in het elektronische patiëntendossier zijn of haar bevindingen noteert. Verder worden laboratoriumuitslagen, vragenlijsten, röntgenfoto's, scans, medicatievoorschriften, bijwerkingen digitaal opgeslagen. Dat levert een enorme hoeveelheid informatie op over heel veel patiënten.

Deze zorggegevens worden nu in rap tempo ontsloten en beschikbaar gesteld voor onderzoek. Veel mensen hebben hier

heel hoge verwachtingen van: de big data hype. Ik ben als statisticus betrokken bij verschillende projecten waar routine zorgdata worden gebruikt in onderzoek. En ik moet helaas deze hoge verwachtingen toch enigszins temperen. Allereerst is het een enorme klus om deze data zo klaar te maken dat je er überhaupt onderzoek mee kunt uitvoeren. Daar gaat vele maanden werk in zitten. De juiste patiënten moeten geselecteerd worden, met het juiste beginpunt en de juiste follow-up. Gegevens komen uit verschillende bronnen die gekoppeld moeten worden. Er zitten invoerfouten in dit soort gegevens. Zo vonden wij in een studie dat er opvallend veel personen in voorkwamen die rond de 80, 90 cm lang waren en daarbij ook nog extreem zwaar waren, rond de 170, 180 kg. Het was meteen duidelijk: regelmatig waren lengte en gewicht bij het invoeren verwisseld. Voor de zorg heeft zo'n invoerfout geen consequenties, de dokter hoeft maar naar de patiënt te kijken om zijn lengte goed te kunnen schatten. Maar zulke fouten kunnen wel de resultaten van analyses vertekenen. Het is een misvatting te denken dat invoerfouten in grote databestanden niet erg zijn omdat ze zouden uitmiddelen. Integendeel, uitschieters kunnen een grote invloed hebben op resultaten van analyses. Bovendien geven meetfouten in confounders of mediatoren systematische vertekening. Goede controle van data is dus nodig.

Gegevens worden alleen goed geregistreerd als er prikkels zijn om dit te doen. Een voorbeeld van dichtbij. Mijn oudste twee kinderen zijn op kamers gegaan toen uitwonende studenten een hogere studiefinanciering kregen dan thuiswonende studenten. Voor hen was het zaak om zich meteen in te schrijven in hun nieuwe woonplaats. In het nieuwe leenstelsel is dit voordeel afgeschaft en is er geen onderscheid meer tussen thuis- en uitwonende studenten. Ongetwijfeld beïnvloedt het nieuwe leenstelsel de keuze van studenten om het ouderlijk huis te verlaten en, erger nog, bij sommigen de keuze om al dan niet een opleiding te volgen.[8] Maar er gebeurt ook iets anders. Juist omdat er geen studiefinanciering meer is, is de prikkel verdwenen om je als student na een verhuizing meteen als inwoner van je nieuwe woonplaats in te schrijven. En zo staat mijn jongste dochter, die alweer een aantal maanden in Delft woont, nog als thuiswonend in Leiden geregistreerd.

Dit voorbeeld illustreert dat het bij hergebruik van data belangrijk is te weten hoe er geregistreerd is, of er verandering in de manier van registreren over de tijd is, of verschillende artsen op dezelfde manier registreren en waar de registratie onvolledig of verkeerd is. Artsen zullen afwijkingen eerder vastleggen dan standaardbevindingen en aanvullende diagnostiek wordt niet bij elke patiënt uitgevoerd, maar alleen bij redelijke verdenking op bepaalde aandoeningen. Als een patiënt een ziekere indruk maakt, worden er meer laboratoriumbepalingen uitgevoerd en zullen er vaker scans gemaakt worden. Verder beïnvloedt de ernst van een aandoening de keuze van behandeling. Bij die keuze spelen ook factoren mee die moeilijk te registreren zijn: hoe presenteert de patiënt zich, maakt hij een vitale indruk, ziet hij bleek, heeft hij voorkeur voor een behandeling? Patiënten worden ook niet op dezelfde manier gevolgd. Mensen met een mildere vorm van een bepaalde ziekte worden minder vaak en minder intensief gecontroleerd.

Analyse van dit soort gegevens is dan ook niet eenvoudig. Er zijn meetfouten, waaronder systematische fouten. Er is sprake van confounding, waarschijnlijk ook door factoren die niet goed geregistreerd zijn. Er is geen gestandaardiseerde follow-up. En er ontbreken veel gegevens, waarbij het feit dat een meting ontbreekt veel informatie kan bevatten over de conditie van een patiënt. Wij kijken op dit moment naar ontbrekende gegevens in dit soort data. Standaard multiple imputatie om de ontbrekende gegevens aan te vullen gaat vaak niet goed, omdat, in statistische termen gesproken, het ontbreken niet at random is. Missende gegevens bij het bouwen van propensity scores, een vaak gebruikte methode om te corrigeren voor confounding, is een van de onderwerpen waar promovendus Jan Choi onderzoek naar doet.

Met de komst van al deze *big data* zijn er vanuit de computerwetenschappen nieuwe vakgebieden ontstaan zoals *data analy-*

tics en *data science*. Maar de problemen waar de *data scientists pur sang*, de epidemiologen en statistici, zich al vele decennia mee bezig houden zijn niet opeens verdwenen. Nadenken over toeval en vertekening blijft nodig. Nadenken over de correcte wijze van gegevens verzamelen, de invloed die het gekozen design heeft op de analyse, het omgaan met meetfouten, met selectie, met confounding, met missende gegevens, met onvolledige follow-up en gecensureerde uitkomsten, met competing risks, met herhaalde metingen en validatie van modellen. De nieuwe technieken vanuit de computerwetenschappen kunnen nuttige aanvullingen zijn op onze analysemethoden, maar ze zijn geen losstaand alternatief en zeker geen panacee. Bovendien is de term *big data* maar relatief. Wanneer er naar specifieke ziekten gekeken wordt, worden aantallen snel kleiner en speelt toevalsvariatie een grote rol.

Naast het beantwoorden van causaliteitsvragen worden grote databestanden ook gebruikt om diagnoses te stellen of prognoses te maken. Er is een enorme ontwikkeling gaande op het gebied van voorspelmethoden. Daaronder vallen *machine learning* en *deep learning* technieken, methoden die hypothesevrij naar patronen in de data zelf kijken. Deze zijn goed te gebruiken voor bijvoorbeeld diagnose op grond van omics data, biopten of beeldmateriaal. Bij klinische voorspelmodellen werken we anders. Daar kiezen we op grond van inhoudelijke kennis mogelijke voorspellers van de uitkomst, het liefst variabelen die eenvoudig te meten zijn, bijvoorbeeld bloeddruk in plaats van een genetische bepaling. We maken verschillende voorspelmodellen. Eerst een model met alleen klinische variabelen. Vervolgens kijken we of biomarkers of genetica daar nog iets aan toevoegen. Regressiemodellen, eventueel met krimptechnieken, worden gebruikt om voorspelmodellen te maken en te valideren. Ik zal niet in detail treden, meer kunt u horen in de oratie van professor Ewout Steyerberg begin volgend jaar.

Regressiemodellen hebben als groot voordeel dat ze begrijpelijk zijn. In een oogopslag is te zien welke variabelen er meegenomen zijn, en hoe deze in het voorspelmodel gewogen worden. Dat het verstandig is om voorspelalgoritmes te begrijpen, volgt uit een recent voorbeeld.[9] Om de kans op overlijden te voorspellen bij patiënten met een longontsteking, werd met machine learning een voorspelmodel gebouwd. Hiermee wilde men bepalen of een patiënt met longontsteking in het ziekenhuis moest worden opgenomen of thuis behandeld kon worden. Tot ieders verbazing voorspelde het algoritme bij patienten die naast de longontsteking ook astma hadden een lager risico op overlijden dan bij patiënten zonder astma. De verklaring voor deze vreemde bevinding werd later gevonden. In de ziekenhuizen waar men het model had ontwikkeld, werden patiënten met astma en longontsteking uit voorzorg al anders behandeld. Ze werden bijvoorbeeld vaker opgenomen op de IC-afdeling. Het voorspelmodel had dit niet meegenomen.

We moeten voorspelmodellen dus begrijpelijk houden en ze goed valideren voordat ze in de zorg gebruikt kunnen worden. Dus geen Black Box-prognoses in de zorg! Verder laat dit voorbeeld zien hoe belangrijk het is om te weten waarvoor een voorspelmodel gebruikt gaat worden. In dit voorbeeld was het doel te beslissen over wel of geen ziekenhuisopname en lag onder dit predictieonderzoek een causale vraag. Van elke patient wil men twee potentiële uitkomsten weten: de uitkomst als de patiënt wel opgenomen wordt en de uitkomst als hij thuis behandeld wordt. Causaliteitsonderzoek en predictieonderzoek zijn dus veel meer verweven, dan menigeen denkt.

De medisch statisticus

De carrièrewebsite CarreerCast.com stelt elk jaar een top 10 van de meest aantrekkelijke banen samen. U mag drie keer raden wat er dit jaar op nummer 1 staat… Statisticus! Helemaal terecht natuurlijk en ik kan daaraan toevoegen, dat het vak van medisch statisticus daar nog bovenuit steekt.

Wat is er eigenlijk nodig om een goede medisch statisticus te zijn? Allereerst een goede wiskundige basis, om de finesses van de statistische methoden te begrijpen en om onderzoek uit te kunnen voeren om methoden te verbeteren. Verder goede pro-

grammeervaardigheden. Maar dat is niet het enige; medisch statisticus zijn omhelst meer dan alleen statistiek. Participeren in medisch onderzoek is een belangrijke taak van ons. Daarvoor is het essentieel om te kunnen communiceren met medisch onderzoekers. Hoe gaat dat? Allereerst is het van wezenlijk belang hierbij (en soms niet makkelijk) om de onderzoeksvraag helder te krijgen. Want alleen dan is het mogelijk om de bijpassende statistische methoden te bepalen, die de onderzoeksvraag gaan beantwoorden. Vervolgens voer je de analyses uit, of wat vaker gebeurt, je legt de onderzoeker uit hoe deze de analyses moet uitvoeren en kijkt met hem mee. Het is dan belangrijk dat de onderzoeker begrijpt wat er gedaan wordt. Tenslotte moeten de resultaten weer terugvertaald worden naar de toepassing zodat het antwoord op de onderzoeksvraag en de klinische implicaties duidelijk zijn. Die combinatie van toepassing en statistiek maakt mijn vak zo boeiend.

Idealiter zijn zowel epidemiologen als statistici ook vanaf het begin bij observationeel onderzoek betrokken. Vanuit verschillende invalshoeken kan er dan naar een onderzoek gekeken worden: wat is de vraagstelling, zijn er confounders, hoe meten we die, is er risico op selectiebias? Nadenken vooraf voorkomt dat essentiële zaken over het hoofd gezien worden, en goed nadenken kan er voor zorgen dat de uiteindelijke statistische analyses een stuk eenvoudiger worden en de resultaten betrouwbaarder. De praktijk is weerbarstiger. Dan wordt de statisticus vaak pas veel later ingeschakeld, bijvoorbeeld nadat de data al verzameld zijn. Of nadat alle analyses al uitgevoerd zijn. Soms zelfs pas als het artikel al geschreven is en de referent nog met vragen over de statistiek komt. Met de research support desk, opgezet door het LUMC research ICT programma en gecoördineerd door doctor Karin van der Pal en doctor Ineke van der Veen, proberen we dit in het LUMC te verbeteren. Onderzoekers in de opstartfase van hun onderzoek (klinisch en niet klinisch), krijgen tijdens een multidisciplinair overleg advies van experts van verschillende methodologische disciplines, zoals de epidemiologie, de statistiek, besliskunde, GRP-commissie en data management.

Negen maanden geleden gaf professor Theo Stijnen in deze zaal zijn afscheidscollege. Ik wil twee zaken waar hij voor pleitte nogmaals benadrukken: reproduceerbaarheid van onderzoek en het opstellen van een rigoureus statistisch analyseplan. In een wereld van *Open Science* wordt hergebruik van bestaande data door instanties zoals subsidiegever ZonMW sterk aangemoedigd. Veel wetenschappelijke tijdschriften vragen om bij publicatie van een artikel de data en metadata beschikbaar te stellen. Dat zorgt voor transparantie, iedereen kan nagaan hoe de resultaten tot stand gekomen zijn. Maar er schuilen risico's in het vrije hergebruik van data. Ze kunnen op een verkeerde manier geanalyseerd worden door mensen die onbekend zijn met het design en de achtergrond van de studie. Men kan data selectief doorspitten op zoek naar speciale verbanden en allerlei secundaire of hypothesevrije analyses uitvoeren. Heel veel toetsen zonder hypothesen vooraf leidt tot toevalstreffers en rare associaties. Als je maar lang genoeg zoekt, vind je altijd wel iets. Dan krijgen we het soort interessante bevindingen als dat liefhebbers van krulfriet intelligenter zijn dan gemiddeld.[10] U kent vast wel meer voorbeelden. Zolang het gaat over krulfriet en bitterballen zijn dit soort resultaten nog wel vermakelijk, maar slecht opgezet of uitgevoerd onderzoek vervuilt de literatuur, vertraagt de wetenschap, en zorgt voor onnodige paniek als resultaten de media halen. Regie over de data is dus nodig, ook bij de zorgdata die voor onderzoek ontsloten wordt. Van wie zijn de data? Wie mogen er onderzoek mee doen? Onder welke voorwaarden? Hoe zorgen we dat de methodologie goed geborgd is? Ik pleit voor gecontroleerd gebruik van medische onderzoeksdata, zoals bijvoorbeeld gebeurd bij de NEO-studie waar de data pas uitgeleverd worden, nadat het volledig protocol met statistisch plan goedgekeurd is.

Een brug tussen theorie en praktijk

Veel medisch statistici voeren ook methodologisch onderzoek uit. Dat kan op verschillende manieren. Men specialiseert zich op een methodologisch deelgebied, een specifieke niche. Of men is een generalistisch statisticus en neemt de onderzoeksvragen die aangedragen worden in samenwerkingsverbanden

met medisch onderzoekers als uitgangspunt voor statistisch onderzoek. Ik heb voor de laatste optie gekozen, statistisch onderzoek gemotiveerd door klinische vragen. Als generalistisch statisticus lever je vaak een heel relevante bijdrage aan medisch onderzoek maar hieruit volgen minder vaak puur methodologische publicaties en minder eerste auteurschappen. Ik hoop dat men daar rekening mee houdt bij de carrièrepaden van jonge statistici. Generalisten zijn belangrijk en blijven nodig, maar als het huidige beleid voor medische onderzoekers rigoureus wordt toegepast op medisch statistici, delven zij het onderspit.

De ontwikkelingen binnen het statistisch onderzoek gaan heel snel, niet alleen in het causale onderzoek maar ook op allerlei andere gebieden. Toch worden veel van de nieuw ontwikkelde statistische methoden zelden gebruikt in medische toepassingen. Daar zijn verschillende redenen voor. Soms zijn ontwikkelde methoden niet bruikbaar in de praktijk. Het is verleidelijk voor statistisch onderzoekers om een model nog net iets ingewikkelder te maken, gewoon omdat het technisch kan. Dat levert met een beetje geluk een mooie publicatie in een statistisch tijdschrift op. Maar op nodeloos ingewikkelde modellen met wiebelige schattingen zit de praktijk niet te wachten. Modellen moeten robuust zijn, liefst zo simpel mogelijk en goed interpreteerbare resultaten opleveren. Soms worden methoden niet gebruikt omdat software ontbreekt, of er is alleen een R package met onbegrijpelijke documentatie beschikbaar dat cryptische foutmeldingen geeft. Ook dat werkt niet. Nadenken over implementatie van ontwikkelde methoden is iets wat vaak vergeten wordt. Het kan anders: promovendus Katerina Papadimitropoulou heeft haar pseudo-IPD-meta-analyse methoden in zowel SAS, SPSS als in R geïmplementeerd in de hoop dat anderen haar methoden ook gaan gebruiken. Tenslotte worden de meeste statistische analyses in medisch onderzoek helemaal niet uitgevoerd door statistici, maar door (bio)medisch onderzoekers, voor wie de statistische literatuur ontoegankelijk is.

Soms vinden statistische methoden wel hun weg naar de medische toepassingen. Dan ontstaat echter weer een ander probleem. Onderzoekers passen deze methoden soms toe zonder rekening te houden met de beperkingen en de onderliggende assumpties. Multiple imputatie van missende waarden is zo'n voorbeeld. Onderzoekers realiseren zich niet altijd dat deze methode naast missing at random, ook veronderstelt dat de onderliggende imputatiemodellen correct zijn. Terug naar mijn vraag aan het begin van mijn verhaal: *Geeft roken een hoger risico op dementie?* Deze vraag kan op twee manieren beantwoord worden. Ja, want roken beschadigt de bloedvaten en slechte bloedvaten kunnen leiden tot dementie. Maar het antwoord kan ook nee zijn. Want roken geeft ook een hoger risico op longkanker en hart- en vaatziekten waardoor rokers al overleden zijn voordat ze dementie kunnen ontwikkelen. Een Cox-regressie zal het eerste antwoord geven, een Fine en Gray-model voor *competing risks* geeft het tweede antwoord, in tegenstelling tot wat veel onderzoekers denken.

Er is dus een grote behoefte aan handvatten om statistische analyses goed uit te voeren, zowel voor statistici als niet-statistici. Hiervoor werkt een grote groep statistici, waar ik ook bij hoor, internationaal samen in een initiatief genaamd STRATOS, om voor veel voorkomende methodologische problemen in observationeel onderzoek, de beschikbare methodologie samen te vatten, om de voor- en nadelen van verschillende methoden te bespreken en om voor voorbeelden met software te zorgen.

Onderwijs

Dit brengt me bij het onderwijs. Het moge uit mijn betoog duidelijk zijn dat goed onderwijs in epidemiologie en statistiek aan onderzoekers en studenten noodzakelijk is. Het is een kunst apart om statistiekonderwijs te geven aan niet-statistici zoals geneeskundestudenten. Belangrijk is om begrip en inzicht te kweken, niet teveel wiskundige formules voor te schotelen en vooral geen trucjes aan te leren. In het derdejaars CAT-project, dat door doctor Arno Roest en mij gecoördineerd wordt, leren studenten om *evidence based* te werken. Hiervoor moeten ze medische artikelen kritisch kunnen be-

schouwen. De meeste artikelen staan tegenwoordig vol statistiek. Het is dus belangrijk dat studenten statistische concepten zoals p-waarden, betrouwbaarheidsintervallen en statistisch significant, begrijpen en op hun waarde kunnen schatten. Die p-waarden, geachte studenten en geachte onderzoekers, daar mag u niet teveel naar kijken. Al te vaak wordt een p-waarde groter dan 0,05 geïnterpreteerd als geen verschil en een p-waarde kleiner dan 0,05 als bewezen verschil. Doe dat niet! Kijk vooral naar de grootte van de effecten met hun onnauwkeurigheid en bedenk of deze klinisch relevant zijn.

Op het zelf uitvoeren van onderzoek bereiden we de studenten voor in het blok Praktische Onderzoeksvaardigheden. Hier behandelen we alle facetten van onderzoek doen. Van het bedenken van een onderzoeksvraag, het schrijven van een studieprotocol, het uitzoeken en uitvoeren van de correcte statistische methoden en het interpreteren van de resultaten. Dat gaat allang niet meer in toga met baret en een lang verhaal, wij proberen studenten actief te laten leren met nieuwe didactische middelen. We hebben net weer een nieuwe e-learning over ethiek en privacy ontwikkeld.

Er is een groot tekort aan statistici. Met het LUMC, het Mathematisch Instituut, de Faculteit Sociale Wetenschappen en de Universiteit Wageningen is er samengewerkt om dit te veranderen. Dat heeft geleid tot een zelfstandige interfacultaire master Statistical Science, waar studenten vanuit verschillende bachelors kunnen instromen. Verschillende biomedische wetenschappen en geneeskundestudenten hebben dit al gedaan. Met alle nieuwe ontwikkelingen in het medisch onderzoek zal de behoefte aan goed opgeleide statistici alleen maar groeien. Ik hoop dat ik met mijn oratie heb laten zien hoeveel uitdagingen er liggen en hoe leuk het is om hieraan te werken.

Een laatste opmerking over onderwijs. Nadenken over de onderwerpen die belangrijk zijn om over te dragen, het geven van onderwijs en de interactie met studenten is niet alleen inspirerend om te doen, maar je leert er als wetenschapper ook zelf veel van.

Dankwoord

Ik wil eindigen met woorden van dank. Mijnheer de Rector Magnificus, leden van het College van Bestuur van de Universiteit Leiden, leden van de Raad van Bestuur van het LUMC: Ik dank u voor het in mij gestelde vertrouwen.

In generaties voor mij was studeren voor velen, vooral voor vrouwen, niet weggelegd. Ik prijs mij gelukkig dat dat ik mijn talenten wel heb kunnen ontplooien, en voel het als een groot voorrecht dat ik u allen hier vandaag mag toespreken. Ik wil iedereen danken die mij gesteund heeft en een bijdrage heeft geleverd aan het feit dat ik hier vandaag sta.

Veel ben ik verschuldigd aan mijn promotor professor Hans van Houwelingen. Hans, van jou heb ik het vak medisch statisticus geleerd. Jij hebt de gave om relevante methodologische problemen te identificeren, en ze wiskundig gefundeerd aan te pakken.

Professor Theo Stijnen, fijn dat ik altijd bij je terecht kon en kan wanneer ik vragen heb, ook nog na je pensioen. Je gedegenheid en je betrokkenheid waardeer ik zeer.

Professor Ewout Steyerberg, een nieuw afdelingshoofd bij de Medische Statistiek met veel plannen. Ik doe daar graag aan mee.

Professoren Ronald Brand, Hein Putter, Jelle Goeman, en alle verdere collega's van de Medische Statistiek, ik wil jullie bedanken voor de statistische discussies, samenwerking en gezamenlijk congresbezoek. Sommigen van jullie ken ik al vele jaren, jullie zijn mij dierbaar.

Vanaf 2007 heb ik een deelaanstelling bij de afdeling Klinische Epidemiologie. Professor Frits Rosendaal, bedankt dat je mij je afdeling hebt binnengehaald, en bedankt voor je betrokkenheid en je vertrouwen in mij.

Professor Jan Vandenbroucke, door jou heb ik de epidemiologische manier van denken over observationeel onderzoek ontdekt. Ik heb daar veel van geleerd.

Professoren Suzanne Cannegieter, Anske van der Bom, Friedo Dekker, Olaf Dekkers en alle verdere medewerkers van de afdeling Klinische Epidemiologie, bedankt voor de goede sfeer, voor de discussies over causaliteit en het feit dat ik vaak nauw betrokken word bij jullie onderzoek, met veel uitdagende methodologische vragen. Yvonne en Tamara, bedankt voor de hulp bij de voorbereiding van vandaag.

Alle onderzoekers bij wiens projecten ik betrokken ben of ben geweest, wil ik bedanken voor alle samenwerking. Ik heb het altijd bijzonder prettig gevonden om een bijdrage te kunnen leveren aan het oplossen van de vele relevante medische vraagstukken waar jullie aan werken.

Iedereen betrokken bij het onderwijs, het is fijn om samen goed onderwijs te ontwikkelen en te geven. Binnen het LUMC valt het mij steeds weer op hoeveel mensen zich enthousiast inzetten voor het onderwijs.

De STRATOS causal inference groep, waarvan professor Els Goetghebeur vandaag hier aanwezig is. Werken met vrouwelijke statistici aan causaliteit is inspirerend en bijzonder plezierig.

Tot slot mijn vrienden en familie. Er is meer dan alleen werk, jullie zijn belangrijk voor mij en ik ben blij dat jullie vandaag hier aanwezig zijn. Ik wil mijn twee zussen Liesbeth en Marijke en mijn schoonmoeder Nel met name bedanken. Mijn ouders zijn allebei helaas overleden en maken dit niet mee.

Lieve Emiel, Hanneke en Marjolein, alle drie studerend, jullie weten van veel onderwerpen veel meer dan ik. De toekomst ligt voor jullie open. Ik heb er vertrouwen in en ben ontzettend trots op jullie .

Lieve Jan Adriaan, je bent al vele jaren mijn steun en toeverlaat en nog veel meer. Heel erg bedankt daarvoor.

Ik heb gezegd

Referenties

1 Boers KE, Vijgen SMC, Bijlenga D, Van der Post JAM, Bekedam DJ, Kwee A et al. Induction versus expectant monitoring for intrauterine growth restriction at term: randomised equivalence trial (DIGITAT). British Medical Journal. 2010;341.

2 De Mutsert R, Den Heijer M, Rabelink TJ, Smit JWA, Romijn JA, Jukema JW et al. The Netherlands Epidemiology of Obesity (NEO) study: study design and data collection. European Journal of Epidemiology. 2013; 28(6): 513-23.

3 Berkson J. Limitations of the Application of Fourfold Table Analysis to Hospital Data. Biometrics Bulletin. 1946; 2(3): 47-53.

4 Hernández-Díaz S, Schisterman EF, Hernán MA. The birth-weight 'paradox' uncovered? Am J Epidemiol 2006; 164: 1115–20.

5 Boef AGC, Le Cessie S, Dekkers OM. Mendelian Randomization Studies in the Elderly. Epidemiology. 2015; 26(2): E15-E6.

6 Le Cessie S. Bias Formulas for Estimating Direct and Indirect Effects When Unmeasured Confounding Is Present. Epidemiology. 2016; 27(1): 125-32.

7 Le Cessie S, Debeij J, Rosendaal FR, Cannegieter SC, Vandenbroucke JP. Quantification of Bias in Direct Effects Estimates Due to Different Types of Measurement Error in the Mediator. Epidemiology. 2012; 23(4): 551-60.

8 Van den Broek A, Wartenberg F, Bendig-Jacobs J, Tholen R,Duysak S, Nooij J (2017). Monitor Beleidsmaatregelen 2016-2017: Studiekeuze, studiegedrag en leengedrag in relatie tot beleidsmaatregelen in het hoger onderwijs, 2006-2016. Nijmegen: ResearchNed, in opdracht van het ministerie van Onderwijs, Cultuur en Wetenschap.

9 Cabitza F, Rasoini R, Gensini GF. Unintended Consequences of Machine Learning in Medicine. Jama-Journal of the American Medical Association. 2017; 318(6): 517-8.

10 Kosinski M, Stillwell D, Graepel T. Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences of the United States of America. 2013; 110(15): 5802-5.

Prof.dr. Saskia le Cessie

1987	Doctoraal Wiskunde, Rijksuniversiteit Utrecht
1991	Promotie Rijksuniversiteit Leiden (Model building techniques for logistic regression, with applications to medical data).
1991-1992	Assistant Professor, Department of Statistics, State University of New York at Buffalo, Buffalo NY, USA.
1992-1993	Assistant Professor, Department of Biostatistics, University of Rochester, Rochester NY, USA
1993-heden	Staflid afdeling Medische statistiek en Bio-Informatica, LUMC, Leiden.
2007-heden	Staflid afdeling Klinische Epidemiologie, LUMC, Leiden
2016	Hoogleraar Statistische Methoden in Observationeel (Klinisch) Epidemiologisch Onderzoek, LUMC Leiden.

Veel vragen over oorzaak en gevolg worden met observationeel epidemiologisch onderzoek bestudeerd. Het is een uitdaging om in dit soort onderzoek met redelijke betrouwbaarheid uitspraken te doen over mogelijke causaliteit. Causale statistische methoden kunnen onder bepaalde voorwaarden hierbij helpen.

Steeds vaker worden routinematig verkregen zorgdata voor onderzoek gebruikt. Deze data zijn niet speciaal voor onderzoek gegenereerd en lastig te analyseren. Om uit dit soort gegevens zinvolle conclusies te kunnen trekken is een gedegen kennis van de epidemiologie en statistiek noodzakelijk.

Universiteit Leiden