Prof.dr. J.J. Houwing-Duistermaat

Statistiek, genetica en epidemiologie

Over de zoektocht in ons DNA naar causale varianten

Universiteit Leiden

Statistiek, genetica en epidemiologie
Over de zoektocht in ons DNA naar causale varianten

Oratie uitgesproken door

Prof.dr. J.J. Houwing-Duistermaat

bij de aanvaarding van het ambt van hoogleraar in de

Genetische Statistiek

aan de Universiteit Leiden

op vrijdag 24 juni 2011

Universiteit Leiden

Mijnheer de Rector Magnificus, zeer gewaardeerde toehoorders.

We gaan terug naar het begin van de 20ste eeuw, naar de tijd na Darwin en Mendel. U kent waarschijnlijk de eerste wet van Mendel nog wel, van de biologieles op de middelbare school. Deze luidt als volgt: elke cel heeft twee kopieën van een gen, ook wel allelen genoemd, behalve de voortplantingscellen, die hebben één allel. Zo geldt voor bepaalde rode en witte bloemen dat twee rode allelen een rode bloemkleur geeft. Twee witte allelen geven een witte kleur, en een rood en een wit allel geven een rode kleur want rood is dominant over wit. Het was aan het begin van de vorige eeuw duidelijk dat de eerste wet van Mendel m.b.t. het overerven van categorische eigenschappen, zoals kleur, heel breed gold. Wat men toen nog niet goed begreep, was hoe continue eigenschappen overerfden die een heel spectrum van waarden aan kunnen nemen. Een voorbeeld van zo'n continue variabele is lichaamslengte. Lichaamslengte correleert binnen families. Immers lange ouders krijgen lange kinderen en korte ouders krijgen korte kinderen. Welk mechanisme was hiervoor verantwoordelijk? Kon de eerste wet van Mendel dit ook verklaren?
Yule en anderen waren ervan overtuigd dat correlatie van continue variabelen binnen families ook verklaard kon worden met de eerste wet van Mendel. Zij dachten dat de variatie van een continue variabele veroorzaakt wordt door een groot aantal allelen, die elk een klein effect hebben op de uitkomst. Anderen, o.a. Karl Pearson, waren het daar niet mee eens. Het was sir Ronald Fisher die in 1918[1] bewees dat Yule gelijk had: hij splitste de variatie van een uitkomstvariabele in erfelijke en niet-erfelijke componenten. Wanneer de erfelijke component uit een groot aantal genen bestaat en de allelen voor deze genen overerven volgens de eerste wet van Mendel, kon de correlatie binnen families, van bijvoorbeeld lichaamslengte, worden verklaard. Met dit en later werk is Fisher van grote betekenis geweest voor de genetica. Het idee van opsplitsen van variatie in verschillende componenten leidde ook tot de ontwikkeling van een statistische methode: de variantie-analyse. Fisher, die wiskunde en natuurkunde in Cambridge had gestudeerd, kan als één van de grondleggers van de mathematische statistiek worden beschouwd. We kunnen wel stellen dat het vakgebied genetische statistiek met het werk van Fisher bijna 100 jaar geleden is ontstaan.
Vanmiddag wil ik u meer over de geschiedenis van de genetische statistiek vertellen. Vooral over de recente geschiedenis na 1980. Ook wil ik u de komende 45 minuten enthousiast maken voor mijn vakgebied. Dit vakgebied is allesbehalve saai, want onderzoek in de genetische statistiek vraagt om een multidisciplinaire aanpak. Een toegepast statisticus moet een brede wetenschappelijke interesse hebben: wij werken samen met epidemiologen, clinici, genetici, biologen, psychologen en bio-informatici. Als toegepast statisticus moet ik wat weten over genetica, over de biologie van oud worden, over de samenhang tussen genen en omgeving in het ontstaan van extreme verlegenheid, over hoe schade door reumatoïde artritis ontwikkelt over de tijd en over de wisselwerking tussen allergie en infectieziekten. Tenslotte wil ik vanmiddag aantonen dat wetenschappelijk onderzoek op het gebied van de medische statistiek van groot belang is voor medisch onderzoek. Immers, de vele gegevens die in epidemiologische studies gemeten en verzameld worden, behoren met correcte en efficiënte statistische methoden geanalyseerd te worden.

In de jaren tachtig van de vorige eeuw groeide het vakgebied genetische statistiek omdat steeds meer studies genetische gegevens bevatten. Men was zich gaan realiseren dat veel ziekten zoals kanker, suikerziekte en hart- en vaatziekten erfelijke componenten hebben en men kon ook steeds beter genetische markers op het DNA meten. Een nieuw vakgebied is daarbij ontstaan: de genetische epidemiologie. Deze is enerzijds uit de epidemiologie en anderzijds uit genetische vakken, zoals de medische en de populatiegenetica, voortgekomen. De epidemiologie bestudeert de samenhang tussen omgevingsfactoren en ziekten op populatieniveau. Bekende voorbeelden zijn het effect van overgewicht op

suikerziekte, en het effect van het dagelijks drinken van een glaasje wijn op hart- en vaatziekten. Genetische epidemiologie houdt zich bezig met de rol van genetische factoren bij deze ziekten. Daarnaast bestudeert en beschrijft genetische epidemiologie het fenomeen dat deze ziekten vaker voorkomen binnen dezelfde families, ook wel clustering genoemd. Dat dit zo is, weet ik van mijn vader. Toen hij ongeveer 25 jaar huisarts was in Balkbrug merkte hij op dat kinderen vaak dezelfde ziekten kregen als hun ouders 25 jaar eerder.
De familiale component van ziekten en andere eigenschappen bestaat uit een combinatie van genetische en omgevingsfactoren. Voorbeelden van omgevingsfactoren die clustering binnen families kunnen veroorzaken, zijn: de levensstijl bij langlevendheid, de mate van hygiëne bij infectieziekten en de manier van opvoeden bij extreme verlegenheid. Wanneer meerdere genetische en omgevingsfactoren een rol spelen bij een ziekte, ontsporing of een eigenschap, spreken we van een complexe uitkomst. In deze rede zal ik het vooral hebben over de zoektocht naar deze genetische factoren.

Eerst wil ik u wat meer vertellen over ons DNA. Het DNA is verdeeld over 23 paar chromosomen. De genetische code wordt gevormd door een lange rij van letters. In deze lange rij komen slechts vier verschillende letters voor. In plaats van letters spreek ik nu even over vier kleuren. We kunnen ons dan een chromosoom voorstellen als een ketting van rode, gele, groene en blauwe kralen. Voor elk paar chromosoom hebben we zo'n kralenketting van onze moeder en één van onze vader gekregen. Er is grote overeenkomst tussen deze kralenkettingen. Op de meeste plekjes hebben ze dezelfde kleur kraal. Wanneer we een plekje op een ketting bekijken dan heeft elke ketting daar bijvoorbeeld een rode kraal, en op een ander plekje heeft elke ketting bijvoorbeeld een blauwe kraal. Het totaal aantal kralen (of letters) over alle chromosomen bedraagt drie miljard. Op ongeveer drie miljoen plekjes zijn de kleuren op de kralenkettingen niet altijd identiek. Meestal zijn er op zo'n plekje met variatie slechts twee kleuren mogelijk. Dus de kraal is bijvoorbeeld rood of geel. Op een ander plekje met variatie kan het bijvoorbeeld om de kleuren rood en groen gaan. Het grootste deel van deze drie miljoen plekjes heeft niets met ziekte te maken. Een gedeelte van deze plekjes wordt gemeten om het DNA van mensen te karakteriseren en hiermee de causale varianten, die wel een rol spelen bij complexe ziekten en eigenschappen, op te sporen.

Om de genetica van complexe ziekten te ontrafelen, stelde Neil Risch in 1990[2] in een reeks van drie artikelen voor om koppelingsonderzoek in paren aangedane, of zieke, familieleden uit te gaan voeren. Deze aangedane paren worden meestal gevormd door twee broers, door twee zussen of door een broer en een zus. Ik noem ze in de rest van deze rede voor het gemak broer\zus paren. Afhankelijk van de grootte van het effect van de genetische variant op de ziekte die wordt bestudeerd, zouden 200 à 300 aangedane broer\zus paren genoeg moeten zijn om een koppeling te kunnen detecteren. Voor de statistische analyse van deze gegevens introduceerde Kruglyak in 1996[3] de "non parametrische linkage" methode, ook wel de NPL-methode genoemd. Bij deze methode zoekt men gebieden op het DNA die binnen de paren gekoppeld zijn met de ziekte die wordt bestudeerd. Hoe zit dit nu precies? Laten we ons eerst beperken tot 1 paar chromosomen. Van dit paar hebben we een chromosoom van onze moeder gekregen en een chromosoom van onze vader. Het chromosoom van de moeder bevat DNA dat oorspronkelijk van haar moeder komt, en DNA dat oorspronkelijk van haar vader komt. Zo bevat het chromosoom dat van de vader komt DNA wat oorspronkelijk van zijn moeder en zijn vader komt. We hebben dus te maken met DNA van vier grootouders. Stelt u zich eens voor dat het chromosoom dat de grootmoeder aan de moeder doorgeeft een reep witte chocola is, en dat het chromosoom dat de grootvader aan de moeder geeft een reep pure chocola is. Op beide repen staan de letters Droste. Nu maakt de moeder een nieuwe reep chocola voor haar kind door de chocola te mengen in een reep, die afwisselend pure en witte chocola bevat. Dit mengen is wel zo gegaan dat de reep precies even

lang is en dat er nog steeds Droste staat. Er worden stukjes chocola uitgewisseld, en de reep chocola voor het kind bevat dus afwisselend witte chocola van de grootmoeder en pure chocola van de grootvader aan moeders kant. Van de vader krijgt het kind ook zo'n mengsel. Laten we zeggen dat hierbij een mengsel wordt doorgegeven van melkchocolade repen met nootjes en zonder nootjes. De stukjes met nootjes kwamen van de moeder van de vader en zonder nootjes kwamen van de vader van de vader.

Een kind heeft dus op elk plekje op het paar repen een stukje chocola van de moeder en een stukje chocola van de vader. Het wordt interessant wanneer we nagaan hoe met de overerving van een specifiek stukje DNA bij een broer\zus paar zit. Laten we ons concentreren op het eerste stukje op de reep: de D van Droste. Het paar D's kan voor het eerste kind witte chocola en melkchocola met nootjes zijn. Het tweede kind kan hier pure chocola en melkchocola zonder nootjes hebben. In een dergelijk geval hebben de kinderen de D's van de vier verschillende grootouders gekregen. Het kan ook zo zijn dat het tweede kind een witte D heeft en een melk zonder nootjes D. Dan hebben de twee kinderen dus chocola van drie grootouders gekregen. Een stukje, namelijk de witte D, hebben ze van dezelfde grootouder gekregen. Tenslotte kunnen de twee kinderen de D's van precies dezelfde grootouders hebben gekregen.
Wanneer we dus de afkomst van de vier D's van een broer\zus paar bekijken, kunnen die van twee, drie of vier grootouders komen. Hoe minder grootouders D's via de ouders aan de twee kinderen doorgeven, hoe meer de kinderen op elkaar lijken wat betreft de soort chocola op die plek. Gemiddeld genomen zullen broer\zus paren de chocola op een specifiek stukje op de reep van drie grootouders krijgen. Er is echter variatie, en deze variatie kunnen we gebruiken om genetische varianten te vinden die geassocieerd zijn met ziekten en andere uitkomstvariabelen.
Stelt u zich nu voor dat er in sommige repen chocola, als foutje, een rozijntje bij de letter R zit. Wanneer dit foutje op een chromosoom zit, hebben mensen een grotere kans om ziek te worden. Wanneer we nu aangedane broer\zus paren selecteren, zal een groot gedeelte van deze paren allebei een foutje op het chromosoom hebben, in dit voorbeeld vertegenwoordigd door een rozijntje bij de letter R. Dat rozijntje heeft het broer\zus paar waarschijnlijk van dezelfde ouder gekregen en daarmee hebben ze hoogstwaarschijnlijk de hele letter R met het rozijntje gekregen. Dat betekent dus dat de vier R's van het paar afstammen van twee of drie grootouders. We zeggen dat dit stukje DNA gekoppeld is met de ziekte. Door op zoek te gaan naar die gebieden op het DNA die gekoppeld zijn met de ziekte, kunnen we een foutje op één van de chromosomen opsporen.

Om nu voor een broer\zus paar op ieder plekje op de chromosomen vast te stellen of het DNA van twee, drie of vier grootouders afkomstig is, hoeven we niet alle drie miljard letters op het DNA te meten. Het is voldoende om verspreid over de chromosomen ongeveer 10.000 plekjes met variatie te meten. Hiermee kunnen we voor een broer\zus paar voor ieder plekje op de chromosomen met voldoende zekerheid bepalen van hoeveel grootouders het DNA afkomstig is. Wanneer we dit voor alle paren hebben gedaan, kunnen we op zoek gaan naar gebiedjes waar het genetisch materiaal in aangedane broer\zus paren vaak van slechts twee of drie grootouders afkomstig is. Om na te gaan of een dergelijk stukje DNA statistisch significant gekoppeld is met de ziekte, passen we de eerder genoemde NPL-toets van Kruglyak toe.

Voor continue uitkomsten, zoals lichaamslengte gemeten in broer\zus paren was al in de jaren 70 een statistische toets ontwikkeld door Haseman en Elston.[4] Mijn collega's Hein Putter en Hans van Houwelingen plaatsten deze toets in de statistische theorie[5] door te laten zien dat deze toets een bijzondere versie is van de score-toets in een variantiecomponenten model. In dit model wordt, à la Fisher, de variatie opgesplitst in verschillende componenten. Wanneer we nu een truc uithalen en in plaats van de continue

eigenschap de genetische markers modeleren, krijgen we als toets voor koppeling een gewogen NPL-toets. De gewichten van de NPL-toets hangen af van de waarden die het broer\zus paar hebben voor deze uitkomst. Zo krijgen de meest informatieve broer\zus paren het meeste gewicht in de analyse. Wanneer extreem concordante paren ook dezelfde genetische varianten hebben en de extreem discordante paren juist verschillende genetische varianten hebben, is er sprake van koppeling.

Voor de Leiden Lang Leven Studie van Eline Slagboom en Rudi Westendorp hebben we de NPL-toets voor aangedane broer\zus paren toegepast, op genetische gegevens van 420 broer\zus paren boven de negentig jaar. Aangedaan is in deze studie trouwens positief, namelijk ouder dan negentig jaar worden. Het doel van Leiden Lang Leven Studie is het vinden van genetische en biologische factoren die een rol spelen bij erg oud worden, ook wel langlevendheid genoemd.

Er is van alles gemeten aan 90-plussers die nog in leven zijn. De uitkomst waarin we echter geïnteresseerd zijn; namelijk de leeftijd waarop zij sterven, kennen we niet. Het liefst zouden we dat natuurlijk wel willen weten, maar aan de personen waarvan we het wel weten, valt weinig meer te meten. Zij zijn immers al dood. We noemen de leeftijden van de 90-plussers in de studie gecensureerd, omdat we alleen maar weten dat hun uiteindelijke leeftijd groter zal zijn dan hun huidige.
De chromosomen van de langlevende broer\zus paren werden in kaart gebracht door 10.000 genetische markers te meten. Marian Beekman berekende voor alle paren op elk plekje van hoeveel grootouders het DNA kwam en paste de NPL-toets toe. In deze studie zitten echter deelnemers van 90, maar ook van 104 jaar. Honderd vier jaar worden is wel heel bijzonder. Negentig jaar is ook een mooie leeftijd, maar wel minder bijzonder. We kennen allemaal wel iemand in onze omgeving die minstens 90 jaar oud is geworden. Wanneer er een gen bestaat met een langlevenallel dan moeten in ieder geval de extreem oude paren dit langlevenallel hebben. De jongere paren hoeven niet perse allemaal een dergelijk allel te hebben. Eline Slagboom wilde daarom in de statistische analyse meer gewicht geven aan de extreem oude paren ten opzichte van de wat jongere broer\zus paren. We onderzochten met welk statistisch model we dit het beste voor elkaar konden krijgen, en hebben een toets voor koppeling tussen genetische factoren en gecensureerde uitkomsten afgeleid. Deze toets bleek ook een gewogen NPL-toets te zijn. Andrea Callegaro[6] heeft in zijn proefschrift een aantal van dit soort gewogen NPL-toetsen afgeleid en beschreven.
We pasten de NPL-toets voor gecensureerde gegevens met succes toe in de Leiden Lang Leven Studie en in een grote Europese studie met hetzelfde design, maar dan met meer dan 2.000 broer\zus paren. Voor langlevendheid hebben we nu in totaal wel zes significante regio's op verschillende chromosomen geïdentificeerd. In deze gebieden liggen waarschijnlijk genen die een rol spelen bij langlevenheid. Dit is een heel mooi resultaat voor het genetisch onderzoek naar langlevendheid, aangezien in de wereld andere genetische studies naar langlevendheid nog niet veel hebben opgeleverd. Een Leidse studie waarbij de koppelingsmethode leidde tot het identificeren van een genetische variant die het risico op osteoartrose verhoogt, is de GARP studie van Ingrid Meulenbelt.[7]

Eind jaren negentig was men toch teleurgesteld in de koppelingsmethode. Men had verwacht dat de genetica van complexe ziekten snel ontrafeld zou zijn, maar de verantwoordelijke genetische factoren bleken moeilijker te vinden dan gedacht. Eén van de oorzaken was dat de studies vaak niet genoeg broer\zus paren bevatten om met voldoende zekerheid koppeling vast te kunnen stellen. De studies waren te klein, omdat men de invloed van een genetische factor op de uitkomst had overschat. Het bleek dat niet één factor verantwoordelijk was, maar dat er meer genetische factoren zijn die elk slechts een kleine bijdrage aan de ziekte leveren. Risch en Meringankas[8] berekenden in 1997 dat in plaats van de gebruikelijke 200 à 300 paren eigenlijk meer dan 2.000 paren nodig waren. Ze stelden in hetzelfde artikel ook een

alternatieve aanpak voor: de genoombrede associatiestudie. Nieuwe meettechnieken moesten het mogelijk maken om circa 500.000 plekjes met variatie verspreid over alle chromosomen te bepalen. Bij deze aanpak zou 1.000 à 2.000 patiënten en controles voldoende moeten zijn om de met ziekte geassocieerde genetische factoren op te kunnen sporen. In 2007 verschenen de eerste artikelen van studies met dit soort gegevens.
Het is tegenwoordig gebruikelijk om de 500.000 gemeten plekjes, de genotypen, aan te vullen tot ongeveer 2,5 miljoen genotypen. Dit doet men door gebruik te maken van een referentiepanel met individuen waarvoor we alle variaties kennen. Het op deze manier invullen van missende genotypen noemen we "imputeren". De standaard statistische analyse van gegevens uit genoombrede associatiestudies is vrij eenvoudig. De benodigde toetsen zitten in elk statistisch pakket. Echter, de meeste van deze pakketten kunnen niet automatisch 2,5 miljoen toetsen uitvoeren. Er is wel software ontwikkeld die deze analyses uit kan voeren, maar deze is alleen geschikt voor de allersimpelste analyses. Voor de meer ingewikkelde studies, zoals familie- en longitudinale studies, waarbij personen in de tijd gevolgd worden, moeten we zelf een script schrijven die de code voor de statistische analyse bevat.

Mijn groep is bij veel van dit soort studies nauw betrokken. Zo ontwikkelt Hae Won Uh methoden voor de familiestudies en houdt Roula Tsonaka zich voor Annette van der Helm bezig met een genoombrede associatie-analyse in reumatoïde artritis patiënten die in de tijd worden gevolgd. Om al dit soort analyses te automatiseren, efficiënt uit te kunnen voeren en om ze gebruikersvriendelijk te maken, ontwierp Stefan Böhringer een "pipeline" die genotypen imputeert en C++- of R-scripts met code voor de statistische analyse uitvoert. Snelheid in de analyses wordt verkregen, doordat de "pipeline" er voor zorgt dat de toetsen voor een aantal genetische markers tegelijk worden uitgerekend. Quinta Helmer zorgt hierbij voor de bio-informatische ondersteuning.

Wat betreft het ontwikkelen van nieuwe statistische methodologie voor genoombrede associatiestudies houden we ons vooral bezig op welke wijze we het beste het effect van een groep genetische markers kunnen toetsen, in plaats van de gangbare methode die het effect van een enkele marker per keer toetst. Het toepassen van de nieuwe statistische methodologie is daarmee vaak efficiënter. U kunt dit vergelijken met het beschikbaar hebben van een product in de winkel. Een tekort aan producten in een winkel kan een aantal oorzaken hebben. Zo kan het tekort veroorzaakt worden door een tekort aan grondstoffen, of door het falen van machines in de fabriek, of door problemen met het vervoer van fabrieken naar de winkels toe. Het is efficiënt om eerst na te gaan waar het probleem optreedt, voordat men in detail de oorzaak gaat zoeken. Zo is het ook met de 2,5 miljoen genotypen. We kunnen eerst voor associatie van specifieke biologische processen toetsen. Dat wil zeggen, we toetsen gezamenlijk alle plekjes op genen waarvan we weten dat ze bij dit specifieke proces betrokken zijn. Wanneer een dergelijk biologische proces statistisch significant is, kunnen we naar de geassocieerde genen gaan zoeken.
Wat betreft het succes van genoombrede associatiestudies heeft de geschiedenis zich herhaald. Ook de aantallen van 1.000 à 2.000 patiënten en controles die nodig zijn om genetische varianten voor complexe ziekten te vinden, waren een onderschatting.

Opnieuw bleek de bijdrage van de meeste genetische factoren veel kleiner dan verwacht, en waren daarom de aantallen te klein om genetische associatie te kunnen detecteren. Dat genoombrede associatiestudies toch een succes werden, komt doordat verschillende onderzoeksgroepen zijn gaan samenwerken. Men publiceert tegenwoordig in een keer een grote meta-analyse waarbij alle data, ook van negatieve studies, worden gecombineerd.

De resultaten uit genoombrede associatiestudies hebben veel nieuwe inzichten in onderliggende biologische mechanismen gegeven. Echter, ze verklaren de clustering van de ziekten

binnen families bij lange na niet. René de Vries kreeg een reumafonds project toegekend met als doel de bijdrage van bekende genetische factoren aan de erfelijkheid van ACPA positieve reumatoïde artritis uit te rekenen. Diane van der Woude en ik hebben berekend, dat 70% van de variatie in ACPA positieve reumatoïde artritis verklaard kan worden door erfelijke factoren.[9] Het HLA verklaart ongeveer 20% van deze erfelijkheid en de andere bekende genetische factoren, waarvan de meeste gevonden in genoombrede associatiestudies, verklaren gezamenlijk nog eens bijna 4%. René de Vries noemde zijn project "het begin van het einde", omdat hij veronderstelde er bijna te zijn. Een betere titel is waarschijnlijk: het einde van het begin. De zoektocht die in de jaren 90 van de vorige eeuw begon, heeft ons een stuk op weg geholpen. Veel genetische factoren zijn gevonden. De eerste stap is succesvol gezet, maar de reis is nog lang niet afgelopen. Ontwikkelingen op het gebied van het meten van grote hoeveelheden gegevens in grote epidemiologische studies gaan nog steeds verder. Het is nu mogelijk om alle letters in alle genen of zelfs voor het hele genoom te meten. Hierdoor kunnen varianten die gemist zijn bij de genoombrede associatiestudies, alsnog worden gevonden. Nadeel is echter dat het hier om enorme grote datasets gaat. We begonnen met 10.000 markers in koppelingsonderzoek en meten nu voor een persoon ca. 3 miljard. De groottes van datasets zijn dus geëxplodeerd. Ook wordt de structuur van de datasets complexer.

Er worden steeds meer stofjes gemeten die in vergelijking met de genetische markers dichter bij de processen in menselijke cellen staan. Een voorbeeld hiervan zijn de metabolieten. Met het meten van deze metabolieten hoopt men meer kennis over de biologische processen te krijgen die een rol spelen bij ziekte. Deze kennis kan op zijn beurt weer de zoektocht naar genetische factoren een stap voorwaarts brengen.
Al deze verschillende datasets, ook wel -omics datasets genoemd, worden in studies gemeten omdat biologen verwachten dat informatie over de ziekte in deze gegevens moet zitten. Fisher[10] heeft eens opgemerkt, dat wanneer de bioloog zegt dat er informatie in een observatie zit, het de taak van de statisticus is die er uit te halen. Dit klinkt eenvoudig, maar dit is het meestal niet. De verschillende datasets hebben met elkaar te maken en zouden eigenlijk ook tegelijk geanalyseerd moeten worden. Op dit moment is dat nog niet mogelijk. De statistische methoden moeten hiervoor nog worden ontwikkeld.

Ik heb u verteld over de recente ontwikkelingen waar mijn vakgebied mee te maken heeft. Graag neem ik u nu mee de inhoud in van het vak statistiek. Eén van de leukste onderdelen is toch het afleiden van een formule die eenvoudig is en intuïtief klopt. Het afleiden van een formule kunt u vergelijken met het maken van een puzzel, bijvoorbeeld een sudoko. Velen van u zullen dit populaire puzzeltje wel eens gemaakt hebben en het tevreden gevoel kennen wanneer het laatste getalletje is ingevuld. Het afleiden van een statistische formule vereist natuurlijk de beheersing van meer technieken dan het invullen van een sudokopuzzel. Bovendien weten we van te voren vaak niet zeker, of er wel een formule bestaat. Des te leuker is het als het lukt om een mooie inzichtelijke formule af te leiden, waarmee de data succesvol en efficiënt geanalyseerd kunnen worden.

Hoe ontwikkelen we een nieuwe statistische methode? We moeten eerst het probleem goed begrijpen. Ik had u verteld over het probleem van te weinig producten in de winkel. Om dit probleem op te lossen, moet u eerst bekijken welke mogelijke oorzaken daaraan ten grondslag kunnen liggen. U zult moeten praten met de importeur van benodigde grondstoffen, met de directeur van de fabriek, de transporteur en de winkelier. U dient daarbij het gehele proces in kaart te brengen.

In ons geval moeten we praten met clinici, epidemiologen, biologen, genetici en bio-informatici. Dan pas kunnen we een model maken dat de relatie tussen de uitkomst enerzijds en de covariaten (genetische en omgevingsfactoren) anderzijds beschrijft. Zo'n model bevat parameters, die geschat moeten

worden met behulp van de gegevens. Ook moeten we statistische toetsen afleiden om de significantie van deze parameters te kunnen toetsen. Hiervoor moeten formules afgeleid en geïmplementeerd worden in scripts. Als deze scripts klaar zijn, kunnen we de data analyseren, de parameters schatten en de toetsen uitvoeren. We zijn dan een heel eind, maar nog niet klaar. Elk model heeft namelijk een aantal aannamen en we moeten nagaan wat er gebeurt als deze aannamen niet helemaal kloppen. We doen dit vaak d.m.v. simulatiestudies. We maken gegevens onder een ander model en passen ons nieuwe model toe en gaan dan na of er niet al te grote fouten optreden. We kunnen de nieuwe methode pas vertrouwen als we dit goed hebben onderzocht en de resultaten positief zijn. Hierbij moet natuurlijk worden opgemerkt dat we niet alle mogelijkheden af kunnen gaan. Ook de informatie die we gebruiken, is nooit volledig. Een perfect model is daarom niet mogelijk. Gelukkig hebben we in de statistiek veel ervaring met wat er fout kan gaan. Kennis en ervaring spelen een grote rol bij het interpreteren en presenteren van uitkomsten van een statistisch model.

De enorme datasets die verzameld worden binnen de genetische epidemiologie wekken ook de interesse van "data miners" en bio-informatici. "Data mining" is een vrij nieuwe discipline die ontstaan is uit computerwetenschappen, artificiële intelligentie en statistiek. Het vakgebied houdt zich onder andere bezig met het identificeren van patronen en onregelmatigheden door grote datasets te doorzoeken. Men gebruikt hiervoor eenvoudige statistische methoden. Oorspronkelijk analyseerden "data miners" vooral hele populaties en hoefden ze geen rekening te houden met onzekerheden waar statistici wel mee te maken hebben bij het analyseren van kleine steekproeven. Tegenwoordig rekenen data miners ook aan steekproeven waarbij veel gegevens beschikbaar zijn. Om toch een uitspraak te kunnen doen of een gevonden patroon in de data wel of niet toevallig is, passen zij permutatietechnieken toe.

Nadelen van "data mining" zijn dat het doorzoeken van een dataset een artefact kan opleveren, dat er meestal geen onderliggend model is dat de samenhang tussen uitkomst en risicofactoren beschrijft, en dat er geen rekening gehouden wordt met de correlatie en selectie bij familiestudies en met informatieve uitval in longitudinale studies. Ook het corrigeren voor covariaten zoals geslacht, roken en leeftijd, is vaak niet of slecht mogelijk. Voor de analyse van grote datasets in epidemiologische studies zijn dus statistische methoden nodig. Fabrice Colas en ik proberen de data-miners-aanpak en de voor epidemiologische studies noodzakelijke statistische methoden met elkaar te integreren. We komen zo tot een methode die meerdere datasets tegelijk analyseert, en die in staat is om uit deze datasets de factoren te identificeren die samen de uitkomst beïnvloeden.

Hoewel we een eind op weg zijn, kunnen de genetische factoren die gevonden zijn in genoombrede associatiestudies in patiëntcontrole-onderzoek, de clustering van ziekten binnen families niet verklaren. Eigenlijk hadden we dit wel kunnen verwachten. Om de oorzaken van clustering van ziekten en andere eigenschappen binnen families te verklaren, zijn juist die familiegegevens van groot belang. Ik pleit voor meer aandacht voor het familie-onderzoek om genetische factoren te vinden. Het is waar dat familiestudies vaak te klein zijn om genetische associatie of koppeling te detecteren. Eén van de uitdagingen moet dan ook zijn om statistische methoden te ontwikkelen die gegevens uit verschillende familiestudies kunnen combineren. Zo houdt Bruna Balliu zich bezig met methoden om kleine familiestudies die verschillen in de manier waarop deze families voor de studie geselecteerd zijn, samen te voegen.
Aan de Leidse universiteit wordt veel familie-onderzoek gedaan. Ik noemde al de Leiden Lang Leven Studie. Een ander voorbeeld is de studie van Maria Yazdanbakhsh. Zij volgt families in de tijd om de wisselwerking tussen infectieziekten en allergie te bestuderen.
Nieuwe familiestudies zijn gepland in het interfacultair

onderzoek in het profileringsgebied "Health, prevention and the human life cycle", dat zich richt op alle stadia van de menselijke ontwikkeling. In driegeneratie families zal de samenhang tussen ziekten of ontsporing enerzijds en genetische en omgevingsfactoren anderzijds worden bestudeerd. Het modeleren van de correlatiestructuren in al deze familiestudies is een boeiende statistische uitdaging.

Statistische consultatie

Het wetenschappelijk onderzoek op het gebied van de medische statistiek wordt meestal geïnspireerd door de statistische consultatie. Voor een specifieke vraagstelling of dataset uit het ziekenhuis blijken de beschikbare statistische methoden niet goed genoeg te zijn. Een nieuwe methode is nodig en wordt ontwikkeld. Deze nieuwe statistische methoden worden gepubliceerd in de biostatistische literatuur, waarbij de al gepubliceerde datasets als illustratie worden gebruikt.
Een nadeel van deze gang van zaken is dat voor de klinische vraagstelling niet de meest optimale methode was gebruikt, want deze moest toen nog worden ontwikkeld.
In de genetische epidemiologie volgen de nieuwe datasets en vraagstellingen elkaar in een rap tempo op. Het is in dit vakgebied uitermate belangrijk dat statistici van tevoren nadenken over hoe de toekomstige datasets eruit zullen gaan zien en de benodigde methoden vooraf ontwikkelen. Ook moeten onderzoekers die een nieuwe studie opzetten, van tevoren nagaan wat er allemaal nodig is, voordat het antwoord op de onderzoeksvraag kan worden opgeschreven. Om statistische kennis en nieuwe statistische technieken tijdig beschikbaar te hebben, moeten onderzoekers samen met statistici de benodigde tijd en geld binnen de studie hiervoor reserveren. Hierdoor kunnen de nieuwe methoden op tijd gepubliceerd worden, zodat ze dan gebruikt kunnen worden bij het analyseren van de data en het beantwoorden van medische vraagstellingen.
Een probleem hierbij is dat statistische vakbladen alleen artikelen accepteren met illustraties van de nieuwe methoden aan de hand van echte gegevens. Een goede illustratie toont immers het nut van de methode aan. Wanneer de datasets nog niet beschikbaar zijn, kan het nut van de methoden worden aangetoond met geavanceerde gesimuleerde datasets.
Echter, statistische tijdschriften vinden simulaties meestal niet voldoende. De reden hiervoor is dat methoden heel goed kunnen werken voor gesimuleerde gegevens, maar wanneer de gesimuleerde data niet genoeg lijken op de toekomstige datasets zal de nieuwe methode later misschien helemaal niet nuttig blijken.
Statistische tijdschriften zouden daarom moeten toestaan en zelfs stimuleren om in artikelen gebruik te maken van geavanceerde gesimuleerde datasets, zoals die bijvoorbeeld beschikbaar zijn in de "genetic analysis workshop". Deze datasets moeten zoveel mogelijk lijken op de toekomstige datasets. Het is dan ook van belang dat een multidisciplinair team deze datasets simuleert. Zo wordt het voor statistici aantrekkelijker om nieuwe statistische methoden te ontwikkelen voor toekomstige datasets en kunnen biologen en medici de resultaten verkregen met de nieuwe methoden, gebruiken in hun artikel. Een mooie reclame voor het vakgebied!

Onderwijs

Geachte studenten. Bij veel vakgebieden speelt kansrekening een rol. Het vervelende met kansrekening is dat je beter hier niet te veel op je intuïtie moet afgaan. In het dagelijks leven schatten en interpreteren we kansen vaak verkeerd. Wanneer de hoofdprijs van de postcodeloterij in de straat naast de onze is gevallen, denken we dat we de komende jaren geen lot meer hoeven te kopen. Wanneer we een huis buitendijks bouwen of kopen, dan weten we dat gemiddeld eenmaal per 100 jaar de waterstand in de rivier zo hoog wordt dat het water het huis binnenloopt. Toen dat in Limburg was gebeurd, dachten meerdere bewoners: "wij hebben ons portie gehad". De verontwaardiging was groot toen het jaar daarna het weer gebeurde. Wanneer een moeder van zeven dochters opnieuw in verwachting is, verwachten we dat het nu toch echt wel een jongen zal zijn. Bij deze voorbeelden gaat het

om onafhankelijke kansen. De gebeurtenis "het vallen van de hoofdprijs in een straat verderop" of "het buiten zijn oever treden van een rivier" heeft geen invloed op de uitkomst van een nieuwe gebeurtenis. Uit de wet van Mendel volgt dat de kans op een zoon 50% is en niet afhangt van het aantal zussen dat er al is. Op de middelbare school heb ik kansrekeningsommetjes moeten maken over het trekken met teruglegging van knikkers uit een vaas met zwarte en rode knikkers, of het trekken met teruglegging van chocolaatjes uit een doosje met pure en melkchocolaatjes.
Wat is de kans dat we eerst een melkchocolaatje trekken en na teruglegging weer een melkchocolaatje? Dat tieners kansrekening niet nuttig vinden, kan ik me goed voorstellen. Bij ons in huis worden de bonbons niet eens in het doosje teruggelegd. We vinden lege doosjes en zakjes terug tussen bergen kleren van onze tieners. Het is duidelijk: sommen zonder teruglegging zijn voor hen veel interessanter. Wat betreft de kansen met teruglegging zijn er binnen de genetica veel leuke sommen te maken; Namelijk de kans op een kind met blauwe ogen, of kans op een kind met blond haar. In de colleges kansrekening in de master "Statistical Science" gebruik ik veel voorbeelden uit de genetica.
Geachte onderzoekers in de medische wetenschappen en geneeskunde studenten in het master plus programma. Vanzelfsprekend is één van onze taken het geven van onderwijs over de nieuwe ontwikkelingen op het gebied van het meten van het DNA, de genetica en de statistische methoden. Samen met Bas Heijmans van de afdeling moleculaire epidemiologie verzorgen we twee cursussen: een basis- en een vervolgcursus. We leren de deelnemers zoveel mogelijk zelf te analyseren. We leren ze ook hoe ze meer uit hun gegevens kunnen halen door optimale methoden te kiezen.
Geachte studenten van de master "Statistical Science" in Leiden en de masteropleidingen Biostatistiek en Bio-informatica aan de universiteit van Limburg, in Hasselt, België. Het is overduidelijk dat voor de analyse van gegevens, verzameld binnen medische studies, vaak geavanceerde statistische methoden nodig zijn. Bovendien moeten ook vaak nieuwe statistische methoden worden ontwikkeld. Het is voor de medische wetenschap van levensbelang dat er biostatistici zijn die de benodigde statistische kennis hebben en vaardigheden beheersen. Deze moeten worden opgeleid. Vanwege de grote impact van genetica in de levenswetenschappen en haar geschiedenis, die al met sir Ronald Fisher in het begin van de vorige eeuw begon, hoort het vak genetische statistiek in het kerncurriculum van zo'n opleiding thuis.

Het belang van biostatistiekonderzoek wordt erkend door de UMC's. Bijna alle UMC's hebben minstens één hoogleraar in de biostatistiek. Het is nu onze taak om de biostatistiek uit te dragen, om toekomstige statistische problemen op te zoeken en de methoden die nodig zijn voor medisch onderzoek te ontwikkelen. Om onderzoek te doen, hebben we PhD-studenten en post-docs nodig. Kort geleden hebben we een werkgemeenschap biostatistiek in de UMC's opgericht. We willen samen het statistische onderzoek stimuleren. We willen cursussen voor onze PhD-studenten en post-docs gaan organiseren en we gaan op zoek naar fondsen.
We hebben natuurlijk ook master-studenten nodig om onze PhD-studenten uit te recruteren. Samen met statistici uit de sociale, landbouw, medische en mathematische statistiek zijn we in Leiden begonnen met de master "Statistical Science". U hoort het: aan de toekomst van de medische statistiek wordt gewerkt!

Geachte toehoorders, ik heb u verteld dat in epidemiologische studies steeds weer nieuwe datasets worden verzameld en gemeten. Deze data behoren correct en efficiënt te worden geanalyseerd, d.w.z. alle informatie die in de data zit, moet eruit gehaald worden. Daarvoor zijn vaak nieuwe statistische methoden nodig. Het ontwikkelen van deze methoden, hoewel heel erg leuk, is vaak niet eenvoudig. Investeringen in nieuwe meettechnieken behoren hand in hand te gaan met investeringen in de bijbehorende statistische methoden. Ook pleit ik voor meer familiestudies. Het zijn juist deze studies die ons verder zullen brengen in het identificeren van factoren

die de clustering van ziekten binnen families veroorzaken. Tenslotte heb ik u verteld dat ik in vele keukens kom en daar mag proeven. Ik heb helaas niet over alle keukens kunnen vertellen. Mijn excuses aan de koks van deze keukens. Mijn rede wil ik nu afsluiten met een woord van dank.

College van Bestuur van de Universiteit van Leiden en Raad van Bestuur van het LUMC, ik prijs me gelukkig te mogen werken in dit ziekenhuis en aan deze universiteit. Ik voel me als een vis in het water met al het multidisciplinair onderzoek dat wordt gestimuleerd via themagroepen en profileringsgebieden. Ik dank u voor het in mij gestelde vertrouwen.
Hooggeleerde van Houwelingen, beste Hans. Jij wilde statistische methoden voor de analyse van genetische en familiegegevens ontwikkelen en gaf het onderwerp aan mij. Je hebt mij het vak geleerd en stuurde mij al tijdens mijn AIO-tijd naar verschillende onderzoekers om over hun datasets en onderzoeksvragen te praten. Hooggeleerde van Duijn, beste Cock. Jouw liefde voor de wetenschap heeft aanstekelijk gewerkt. Het was in mijn tijd in Rotterdam dat ik de gedachte liet varen om het moederschap met een simpele baan in de statistische ondersteuning te combineren. Hooggeleerde Stijnen, beste Theo, dank voor het gestelde vertrouwen. We weten elkaar te vinden en ik vind bij jou de steun die ik af en toe nodig heb in mijn ingewikkelde baan. Beste collega's van de afdeling medische statistiek. Door al jullie activiteiten op het gebied van de statistiek wordt mijn creativiteit sterk gestimuleerd. Bart Mertens die alles weet over het analyseren van proteomics en Jelle Goeman die hoofd is van het Bio-informatica Expertise Centrum wil ik hierbij noemen. In het bijzonder wil ik Hae Won Uh, Lies de Kler en Ron Wolterbeek bedanken. Zij zorgen er voor dat ik niet alleen maar statistiek doe. De Genstat groep. Lieve allemaal. Ik vind het geweldig om met jullie een groep te vormen. We hebben een brede kennis aan statistische methoden en hebben altijd weer de wil en energie om nieuwe uitdagingen aan te gaan. Hooggeleerde Slagboom, beste Eline, het is stimulerend om met jouw groep te mogen samenwerken. Ik zou graag meer tijd willen hebben voor de Leiden Lang Leven Studie. Er is zoveel leuks in deze studie om samen te ontwikkelen.
Mijn ouders, Henk en Tineke. Jullie overtuiging dat het allemaal goed komt en jullie vertrouwen dat ik de goede keuzes maak, hebben het mogelijk gemaakt ook werkelijk deze keuzes te maken. Mijn schoonouders Anno en Erna, jullie staan altijd voor ons klaar. Erik-Jan en ik hebben meestal onze zaakjes wel geregeld, maar als er iets mis gaat, zijn jullie er altijd. Familieleden en vrienden. Ik weet het ik werk te hard en heb te weinig tijd voor jullie. De wandelingen die we maken zijn altijd te kort. Ik waardeer het heel erg dat jullie hier zijn. In het bijzonder wil ik Francesca Martella en Li Hsu bedanken. Dear Francesca I had a great and fruitful time in Rome last summer. Dear Li, we share our profession, our struggle to combine being a statistician and a mother and our love for good food, preferable made together in one of our kitchens. I am very pleased that you, Charles, Bryan and Anna are here in Leiden this day.
Tenslotte mijn thuis: Lieve Lise, Krijn en Smilla. Jullie zijn het allerbelangrijkste in mijn leven. Ik ben boven alles jullie moeder. Lise, je kan zulke mooie en ontroerende dingen maken. Krijn en Smilla, ik zou willen dat er meer maandagen en donderdagen in de week zaten. Het kopje thee drinken op die dagen is een gezellige en ontspannende onderbreking van de drukke dag. Lieve Erik-Jan. Jij bent het allerbijzonderste wat mij is overkomen. Al bijna 23 jaar zijn wij samen op pad en gaan wij samen vele uitdagingen aan, ook deze. Zonder jou had ik hier niet gestaan.

Ik heb gezegd.

Literatuur

1 Fisher RA. 1918. The correlation between relatives on the supposition of Mendelian inheritance. Trans Roy Soc Edinb 52: 399-433.

2 Risch N. 1990. Linkage strategies for genetically complex traits, I, II, III. Am J Hum Genet 46: 222-253.

3 Kruglyak L et al. 1996. Parametric and nonparametric linkage analysis: A unified multipoint approach. Am J Hum Genet 58: 1347-1363.

4 Haseman JK and Elston RC. 1972. The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2: 3-19.

5 Putter H et al. 2002. Score test for detecting linkage to quantitative traits. Genet Epidemiol 22: 345-355.

6 Callegaro A. 2010. Using survival data in gene mapping. Doctoral Thesis, Leiden University.

7 Meulenbelt I et al. 2008. Identification of DIO2 as a new susceptibility locus for symptomatic osteoarthritis. Hum Mol Genet 17: 1867-1875.

8 Risch N and Merikangas K. 1996. The future of genetic studies of complex human traits. Science 273: 1516-1517.

9 Van der Woude D et al. 2009. Quantitative heritability of anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis. Arthritis Rheum 60: 916-923.

10 Yates F and Mather K. 1963. Ronald Aylmer Fisher. 1890-1962. Biographical memoirs. The Royal Society 9: 91-129.

Prof.dr. J.J. Houwing-Duistermaat

1986	VWO diploma Meander College, Zwolle
1987	Vrije Hogeschool, Driebergen
1992	Doctorandus in de wiskunde. Universiteit Utrecht
1997	Promotie “Statistical methods for family data” Universiteit Leiden
2006	VIDI, NWO
2010	Benoeming hoogleraar Genetische Statistiek Universiteit Leiden
2011	Oratie “Statistiek, genetica en epidemiologie. Over de zoektocht in ons DNA naar causale varianten” Universiteit Leiden

Universiteit Leiden