HET ZIT IN DE GENEN, TOCH?

Inaugurele rede Mark van de Wiel, uitgesproken op vrijdag 23 mei 2014 in de Aula van de Vrije Universiteit.

[HET ZIT IN DE GENEN, TOCH?]

Mijnheer de Rector, dames en heren,

Met deze oratie aanvaard ik mijn hoogleraarschap in Statistics for Genomics, ofwel Statistiek voor Genoomanalyse. Ik ben verheugd en vereerd dat dit hoogleraarschap is ingebed zowel bij de Faculteit Exacte Wetenschappen als bij het VU medisch centrum. Deze dubbelaanstelling verraadt meteen mijn positie: een spin in het web. Ik hoop deze positie hier te verduidelijken door de verschillende rollen van de statisticus in de wetenschap te belichten. Hierbij concentreer ik me natuurlijk vooral op die tak van de wetenschap die Genomics heet.

Laat ik eerst even de titel van deze rede uitleggen. Het eerste deel zal duidelijk zijn: veel van onze goede en slechte eigenschappen zijn vastgelegd in onze genen of ontstaan tijdens ons leven door spontante mutaties in ons genoom. We hebben blauwe ogen, we blijven lang leven, we worden ziek: de genen spelen vaak een cruciale rol.

In de titel zit de statistiek verborgen achter dat ene woordje met het vraagteken erachter: “toch?”.

[DE SCEPTISCHE STATISTICUS]

De primaire rol van de statisticus in de wetenschap is om vraagtekens te zetten. Vraagtekens achter beweringen van andere wetenschappers. Alleen als de data de twijfel weg kunnen nemen, zijn wij tevreden. In mijn toepassingsgebied, de genomics, heeft men de afgelopen jaren nogal last gehad van een ongebreideld optimisme. Dit betrof dan vooral de voorspellingskracht en causaliteit van metingen aan het genoom. Zo dacht men in de late jaren negentig met die genoommetingen ziektebeloop tot wel 99% accuraat te kunnen voorspellen. Helaas bleek al snel dat men letterlijk door de vele bomen het bos niet meer zag. Dit percentage was vaak op verkeerde statistische methoden ge-ent.

Er wordt wel gezegd: “There are lies, damn lies and statistics”. Ik zou willen zeggen “wrong statistics”. Met goede statistiek lieg je niet, sterker nog: met goede statistiek achterhaal je leugens, of op zijn minst foute conclusies. Een beroemd historisch voorbeeld in de genetica is het experiment van Mendel met de erwtenplanten. Op dit experiment heeft hij zijn erfelijkheidswetten gebaseerd. Statisticus én geneticus Fisher gebruikte Pearson’s chi-kwadraat toets om aan te tonen dat Mendels resultaten te goed om waar te zijn waren. Wetenschappers zijn dus net mensen, en zelfs briljante wetenschappers als Mendel gaan weleens over de schreef. U weet allemaal dat wetenschappelijke integriteit momenteel een veelbesproken onderwerp is. Toch denk ik dat, wanneer het genomisch onderzoek betreft,

wetenschappelijke nalaatbaarheid een veel negatievere invloed heeft op de kwaliteit dan moedwillige fraude. Te vaak publiceren absolute toptijdschriften genomisch onderzoek waarvan de statistiek gewoon slecht is.

Een voorbeeld. Methylatie is een belangrijk proces op het DNA waarmee genen kunnen worden uitgeschakeld. Een artikel in Nature uit 2011 beschrijft methylatie van gebieden op het genoom in stamcellen en normale cellen. Men rapporteert vele gebieden die verschillend gemethyleerd zijn tussen deze twee types cellen. Prachtige plaatjes, krachtige conclusies. Gegeven de huidige interesse in stamcellen begrijpt u het belang van deze resultaten. Het artikel is dan ook al meer dan 600 keer geciteerd. Ik was geïnteresseerd, want ik dacht dat men maar liefst 10 biologische herhalingen had per conditie. En u weet: statistici zijn dol op herhalingen en 10 is veel voor de dure techniek die werd gebruikt, 'bisulphite sequencing'.

Echter wat bleek: de herhalingen waren fictief. Men had het genoom in stukjes opgehakt en 10 achtereenvolgende stukjes als onafhankelijke herhalingen beschouwd. Dus gedaan alsof dát de echte biologische variatie representeert. Dat is zoiets als zeggen dat de variatie tussen zoogdieren bestudeerd kan worden door de variatie tussen mensen te beschouwen. Voor het artikel leidde de sterk onderschatte variatie natuurlijk tot prachtige p-waardes en False Discovery Rates. Deze waren blijkbaar klein genoeg om de referenten te overtuigen. Ik heb het aangekaart bij Nature, waarna een interessante, nog niet afgeronde discussie volgde.

Een zekere scepsis blijft dus noodzakelijk. Want als we 20.000 genen meten, of nog veel meer stukjes van het genoom, dan lijkt er altijd wel eentje tussen te zitten die wat doet. Alleen rigoreuze statistiek zorgt ervoor dat de wetenschappelijke literatuur bespaart blijft van zeer vele fout-positieven. Bovendien wordt er dan niet onnodig veel geld en tijd verspild aan mislukte validatie-experimenten.

[DE COMBINERENDE STATISTICUS]

Maar goed, de essentie van genomics is prachtig. Immers, het is en blijft fantastisch dat men met één experiment simultaan al die 10 tot 100-duizenden stukjes van het genoom kan doormeten. Ik kan u vertellen dat wanneer ik wiskunde-studenten vertel over de ontwikkeling van de genomics in de afgelopen 15 jaar, hun respect voor de moleculaire biologie en de technologie erachter exponentieel toeneemt. Zoals ik net al schetste, heeft de algemene kennis van de statistiek deze stormachtige ontwikkelingen helaas niet kunnen bijhouden. Maar het moet gezegd: gelukkig verschijnen er ook genoeg goede artikelen over genomische data en de toepassingen ervan.

Maar laten we niet eenkenning zijn. Genomics moet wél in perspectief worden gezien van de andere data bronnen die beschikbaar zijn. Gelukkig worden weer steeds vaker ook andere variabelen gebruikt om ziektebeloop te begrijpen en te voorspellen. Hierin schuilt ook een mooie statistische uitdaging: hoe combineren we variabelen met zeer verschillende dimensies in één voorspelling? Zelf houd ik me in het kader van een Europees project bezig met het voorspellen van terugkeer van orale kanker. We zien daar dat de genomische metingen op de primaire tumor te weinig voorspellingskracht hebben. We hopen

door combinatie met gegevens uit MRI of CT scans en allerlei klinische variabelen de voorspellende waarde fors op te krikken.

Dat combineren is leuk, maar vaak is de voorspelling op basis van de standaard beschikbare gegevens voor veel patiënten al heel behoorlijk. In het licht van de toenemende zorgkosten zou je misschien alleen dure extra metingen willen doen voor díe patiënten waarvoor je twijfelt aan de initiële voorspelling. Daarnaast kan het ook zijn dat je die extra metingen in de praktijk liever niet doet, omdat ze niet comfortabel zijn voor de patiënt.

Dus bouwen we ook stapsgewijze predictiemodellen, waarbij je niet alles hoeft te meten aan alle patiënten. Deze hebben we toegepast op de bekende borstkankerdata van het Nederlands Kanker Instituut. Het bleek dat je terugkeer van de tumor net zo goed kunt voorspellen als je maar voor een 1/3 deel van de patiënten de genomische data toevoegt aan de bekende klinische voorspellers. Uiteraard kan het in andere settings ook andersom. Mocht je juist een goedkope en goed gestandaardiseerde genomische test hebben, dan kan deze juist andere variabelen goeddeels vervangen.

[DE SEMI-OBJECTIEVE STATISTICUS]

Het combineren of integreren van verschillende data types, die dan allen beschikbaar zijn voor dezelfde individuen, is inmiddels dus redelijk populair binnen de statistiek. Het combineren van data met voorkennis, of externe informatie, is echter veel minder gemeengoed. Veel statistici verschuilen zich onder het mom van 'objectiviteit' vaak achter de heilige, op zichzelf staande dataset. Vaak negeert men dus grotendeels externe informatie. Eerlijk gezegd denk ik dat de echte redenen vaak luiheid en angst zijn. Men wíl simpelweg niet nadenken over wat relevant is en wat niet. Echter, in deze tijd waar zoveel genomische gegevens publiek beschikbaar zijn, en gegeven de complexiteit van de schattingsproblemen, denk ik niet dat het in de wind slaan van externe informatie de juiste weg is.

Een ander argument om geen externe informatie te gebruiken, is dat die informatie dan niet meer gebruikt kan worden om te valideren. Dat is waar, maar u moet weten dat voor bijna elk gen dat gevonden wordt, wel een mooi biologisch verhaal is dat wordt gestaafd door minstens één studie. Met andere woorden: men zoekt de validatie er wel à posteriori bij. Is dat objectiviteit? Nee, dan denk ik dat het beter is vóór de analyses je aannames en gegevens op tafel te leggen, en te zeggen: hier ga ik het met doen.

In het echte leven maken we continu gebruik van voorkennis om inschattingen te doen. Een voorbeeld uit het casino, een plek waar een statisticus zich natuurlijk thuis voelt. Als iemand veel geld heeft verloren met een nieuw spel in het casino, en ik schat in dat die persoon redelijk intelligent is, dan speel ik liever het spel helemaal niet, dan dat ik eerst duizenden euro's uitgeef aan het spel om vervolgens waarschijnlijk te concluderen dat het voor mij ook niet werkt...

De komende jaren wil ik veel onderzoekstijd investeren in hoe externe informatie op een elegante manier verwerkt kan worden in de analyse van een gegeven genomische dataset. Voor de statistici

onder u zal het niet vreemd klinken dat ik daarvoor zeker ook Bayesiaanse methoden ga gebruiken. Bayesiaanse methoden zijn bij uitstek geschikt: de externe informatie kan verpakt worden in de zogenaamde prior, of à priori verdeling.

Kritiek uit de niet- Bayesiaanse hoek is dat in kleine, laag-dimensionale studies de keuze van de à priori verdeling subjectief is. Welnu, dat is het mooie aan genomische data: we hebben zoveel data van soortgelijke entiteiten, bijvoorbeeld genen, dat het mogelijk is om de à priori verdeling te schatten. Dat noemen we 'Empirical Bayes'. We hebben laten zien dat dit werkt, daar kom ik straks nog op terug. De term semi-objectief laat zich nu verklaren: de te gebruiken bronnen van externe informatie zijn goeddeels keuzes, maar de relevantie van die bronnen wordt bepaald door de data zelf.

Laat ik even terug gaan naar het casino voorbeeld, en ik neem aan dat het nieuwe spel een behendigheidsspel is, zoals bijvoorbeeld Black Jack. Stel nu dat niet alleen ik voor de keuze sta of en hoe vaak ik het spel ga spelen, maar ook u allemaal. Bovendien hebben we allemaal al een ander, soortgelijk spel al heel vaak gespeeld. En helaas: we verloren daar gemiddeld gezien wat.

Nu zouden wij gelijk kunnen besluiten dít spel ook maar niet te spelen. Maar een slimmere strategie is de volgende: wij spelen allemaal het spel een paar keer met geringe inzet, waarna ieder voor zich gaat beslissen of men doorgaat of niet. Waren we toen we dat andere spel speelden allemaal dronken, dan zullen we het nu met dít spel een stuk beter doen, hoop ik. En als 'dronken zijn' redelijkerwijs impliceert dat wij met dat vorige spel zomaar wat deden, dan zullen de data uitwijzen dat de correlatie tussen onze prestaties bij het nieuwe en oude spel zo goed als afwezig is. In dit geval vertellen de data ons dat we weinig gewicht moeten geven aan de externe informatie, in dit geval onze prestaties bij het oude spel. Dan moet ieder voor zich een beslissing nemen op basis van de kleine hoeveelheid gegevens voor het nieuwe spel.

Het wordt echter interessanter als wij bij het vorige spel nuchter waren. En helemaal als de data van beide spelen ons dan vertellen dat er wél een behoorlijke correlatie is tussen de prestaties. Blijkbaar betreft het dan soortgelijke vaardigheiden die nodig zijn om beide spelen relatief goed te kunnen spelen. Let wel: die correlatie kunnen we behoorlijk accuraat schatten, want wij zijn hier, gelukkig, met velen. We kunnen dan veel meer gewicht toekennen aan de externe informatie. Juist omdat voor íeder van ons de data van het nieuwe spel van zeer beperkte omvang zijn, kan dit enorm helpen om betere beslissingen te nemen dan wanneer we deze informatie negeren. Niet voor iedereen zal de beslissing beter zijn, maar gemiddeld gezien vaak wel. En wellicht, maken we als groep zelfs winst. Als wij nu vooraf afspreken dat we die winst delen, dan is iedereen blij.

Bayesiaanse procedures zijn dus zeer geschikt voor het includeren van externe informatie. Ik vind alleen wel dat we ze aan klassieke criteria moet onderwerpen, juist waar het inferentie aangaat. Heeft het 95% credibility interval de juiste afdekking? Schat de Bayesiaanse False Discovery Rate ook echt accuraat de proportie fout-positieven? Kortom: 'Bayes' mag geen excuus zijn om niet na te denken over hoe toevalsvariabelen zich zouden gedragen bij replicatie van het experiment. Het zal voor mij ook geen excuus zijn om niet na te denken over klassieke oplossingen, die vaak minder flexibel, maar wel computationeel efficiënter en dus praktisch aantrekkelijk zijn.

Goed, tijd voor een aantal voorbeelden van hoe we externe informatie zouden kunnen includeren bij de analyse van genomische data.

In het VUmc is één van de speerpunten op het gebied van kankeronderzoek het vinden van genomische markers voor vroeg – diagnostiek. Dat zijn dan stukjes van het genoom die zich in voorfases van kanker al anders gedragen dan in normaal weefsel. Bij die voorfases kunt u bijvoorbeeld denken aan voorloperlesies of adenomen bij respectievelijk baarmoederhals- en dikke darmkanker.

Voor vroeg - diagnostiek wil je een eenvoudige, goedkope en niet al te oncomfortabele test hebben. Dat betekent vaak dat men lichaamsmateriaal zoals uitstrijkjes of bloed moet gebruiken, hetgeen een zeer sterke verdunning is van eventueel aangedaan weefsel. Dus we zoeken vaak naar een heel subtiel signaal, en dan ook nog voor soms wel 500.000 kandidaatmarkers tegelijk. Een heel kleine speld in een grote hooiberg dus. We hebben echter vaak ook genomische data van het aangedane weefsel zelf tot onze beschikking, meestal dan wel van andere individuen. Dan helpt het enorm de relevantie van een mogelijke marker in deze meer uitgesproken weefsels als gewicht mee te nemen voor het subtiele vergelijk waarin je geïnteresseerd bent. Dit kan met een gewogen False Discovery Rate of met een à priori verdeling. Ik heb het uitgeprobeerd voor methylatiedata en het was mooi om te zien dat het werkte: de kleine speld lichtte op in die grote hooiberg.

Een soortgelijke weging kunnen we ook toepassen op predictie, al eerder genoemd en een belangrijk thema in een academisch ziekenhuis natuurlijk. We proberen vragen te beantwoorden als: Kunnen we de genomische data gebruiken om de kans op uitzaaiïngen van een tumor te voorspellen? Of nog ambitieuzer: kunnen we overleving voorspellen met behulp van genomische data?

Bij zulke voorspelproblemen is het niet nodig is om het gewicht van de genomische variabelen van tevoren vast te zetten. Je kunt adaptieve procedures ontwikkelen die de weegfactoren schatten met behulp van een kenmerk van de variabele, zoals op welk chromosoom het ligt. Als dat kenmerk dan niet relevant blijkt, zullen de gewichten ook allemaal dichtbij één liggen en dus, zoals gewenst, weinig effect hebben. Als het kenmerk echter wel relevant is, zullen de gewichten veel variabeler zijn en een duidelijk effect hebben op de predictie. In de praktijk hebben we gezien dat dit principe de accuraatheid van de predictie met enkele procenten kan verhogen. Let wel: elke procent representeert vaak een aantal patiënten, dus iedere procent telt.

Tenslotte moleculaire netwerken. Zoals u wellicht weet, werken genen niet alleen, maar in interactie met elkaar en met vele anderen moleculaire entiteiten zoals eiwitten, microRNAs en ga zo maar door. Een hoog-dimensionaal netwerk is vaak een complexe haarbal. Deze stelt je echter wel in staat om na verstoring van een gen, bijvoorbeeld door een medicijn, de effecten op andere genen te voorspellen.

Statistici houden zich bezig met het schatten van zulke netwerken uit data. Een hels karwei, want die hooiberg waar ik het eerder over had is nu een hooiberg in het kwadraat. Het aantal mogelijke verbindingen tussen p genen is immers van de orde p^2. Maar gelukkig zijn statistici niet de enigen en zeker niet de eersten die hieraan werken: systeembiologen hebben al veel stukjes van die netwerken gedocumenteerd, in ieder geval onder experimentele condities.

Natuurlijk kan zo'n netwerk anders zijn voor ziek weefsel dan voor gezond weefsel. Het zou echter vreemd zijn als een gen met zeer veel connecties onder de ene conditie weinig connecties zou hebben onder de andere conditie. Behalve deze connectiviteit zijn er natuurlijk veel meer eigenschappen van een gen of zelfs een groep genen te kwantificeren in een à priori netwerk. De uitdaging is om deze in een model voor de à priori verdeling te verwerken, en om dat model te schatten met behulp van de data. We zijn er voor dit probleem zeker nog niet uit hoe we dat precies moeten doen en hoe waardevol de externe informatie echt is. Maar goed, wetenschap is vooral leuk als je de antwoorden nog niet weet.

Dit zijn slechts drie voorbeelden, maar het principe is breed toepasbaar in de statistiek voor genoomanalyse. Ik wil benadrukken dat ik in een goede positie ben om te werken aan het ontwikkelen van zulke adaptieve statistische methoden en wel om drie redenen: A) De VUmc omgeving: Het contact met de biomedici is essentieel, want hun kennis heb ik nodig om erachter te komen wat de relevante externe informatie zou kunnen zijn; B) De wiskunde omgeving: Er is in Nederland een toenemende interesse in keuze van à priori verdelingen die leiden tot goede eigenschappen; en C) Het vakgebied leent zich ervoor: In het genomisch onderzoek is men gelukkig al jaren lang verplicht om data publiek te maken bij publikatie. Daarnaast zijn er ook veel databases met allerlei soorten structurele genomische informatie.

[DE WISKUNDIGE STATISTICUS]

Een statisticus die werkt aan genomische data bevindt zich dus op het kruispunt van moleculaire biologie, toegepaste statistiek en wiskunde. Wat is de rol van die laatste ten opzichte van de andere twee? Een wiskundige beschouwt zijn/haar wetenschap als het fundament van zo'n beetje alle wetenschap. Je kunt dus redeneren dat je met voldoende wiskunde- kennis elk statistisch probleem aan kan. Misschien waar, maar dat is een lange, inefficiënte weg als je onvoldoende op de hoogte bent van de beschikbare statistische technieken. Bovendien is naast het oplossen van een probleem het herkennen van interessante problemen minimaal even belangrijk. Alleen door vaak te overleggen met biologen en hun vragen te kunnen koppelen aan beschikbare statistische technieken, zul je in staat zijn te herkennen wanneer die technieken tekort schieten. En zo kom je dus tot uitdagende en relevante statistische onderzoeksvragen.

Een voorbeeld uit de eigen praktijk. Zo'n twee jaar geleden werden we benaderd voor de analyse van RNA sequencing count data. Onze opdracht was om verschillen te vinden tussen vijf gebieden in het brein. De proefopzet was best uitdagend: onvolledig, ongebalanceerd, met metingen aan slechts 7 individuen (het gaat tenslotte om hersenmateriaal), en er was ook sprake van een batch-effect. Maar goed, in eerste instantie dachten we: niets nieuws onder de zon, we gebruiken het populaire R-pakket 'edgeR'. Dit is gebaseerd op de Negatief Binomiale verdeling, het voorziet in robuuste schatters van de overdispersie parameter, en het maakt voor inferentie gebruik van het welbekende GLM-raamwerk. Goed geïmplementeerd en lekker snel.

Maar al snel zagen we de tekortkomingen: we observeerden veel meer nullen dan de Negatief Binomiaal kan verklaren en random effecten konden niet worden geïncludeerd. Bovendien waren de

schatters voor de parameters waarin wij het meest geïnteresseerd waren niet erg robuust. Zie hier de start van een nieuw project: hoe kunnen we dit beter doen?

Dan komt de wiskunde om de hoek kijken. Als mooie en efficiënte taal om methoden te ontwikkelen en te communiceren. Maar juist ook om het inzicht te verkrijgen dat je de hoge dimensie van de data kunt gebruiken om tot betere schattingen te komen, met empirical Bayes methoden zoals ik eerder al aangaf. En tenslotte om de generieke waarde van de methode te verhogen. Wiskundige modellen en de aannames erachter geven anderen de gelegenheid om in te schatten in hoeverre een specifieke methode van toepassing is op hun data, en ze kunnen erop bouwen als aanpassingen nodig zijn.

Voor de RNA sequencing breindata was de wiskunde essentieel om een model te bouwen dat deze data beter beschrijft en voor het ontwikkelen van nieuwe, robuuste schattingsmethoden. Na implementatie resulteerde dit tot een heel veelzijdig en flexibel software pakket. Dit is ook al toegepast op andere data sets zoals microRNA sequencing data voor uitgezaaide dikke darmkanker en high-throughput screens voor het effect van gene silencing in een medicatie context.

[DE ECONOMISCHE STATISTICUS]

Mijn middenpositie is soms handig, want je kunt van meerdere walletjes mee-eten, bijvoorbeeld wanneer het gaat om een hoogleraarschap. Maar, om kort even banaal te zijn, het kan een voordeel én een nadeel zijn als het gaat om geld, beurzen dus. Onze multi-disciplinaire aanpak komt zeker van pas bij Europese beurzen voor grote consortia. We zijn betrokken bij twee van zulke consortia, één op het gebied van stralingsgeïnduceerde kanker en één op het gebied van orale kanker. Het is fijn dat onze toegevoegde expertise hierbij wordt gewaardeerd, maar er is natuurlijk maar beperkt ruimte voor methodologische ontwikkeling. En dan laat ik de bijkomende Europese bureaucratie en de inefficiëntie van grote consortia verder maar even terzijde.

Persoonsgebonden beurzen zijn hét gereedschap om meer fundamenteel methodologisch onderzoek te kunnen doen. Maar helaas, zowel bij de biomedici van Zon-MW als bij de wiskundige geledingen van NWO Exact Wetenschappen behoren wij tot de periferie. Ja, er wordt vaak gepraat over multi-disciplinair onderzoek en de waardering ervoor, maar er wordt niet op afgerekend. Het zou mooi zijn als een bescheiden deel van het budget nadrukkelijk gereserveerd wordt voor gebiedsoverschrijdende voorstellen, ook wanneer het persoonsgebonden beurzen aangaat. Zo voorkomen we dat appels met peren worden vergeleken, en stimuleren we wetenschappers om ook echt multi-disciplinair onderzoek te doen.

[DE MEDISCH STATISTICUS]

Mijn eigen beurs wordt gespekt door het VUmc waar ik met plezier werk in de afdeling Epidemiologie & Biostatistiek. Er wordt wel eens geopperd dat het beter zou zijn om statistici onder te brengen bij andere afdelingen. Ik snap dat dit het gevoel met de data zou kunnen verbeteren en ook zou kunnen

leiden tot intensievere samenwerkingen. Ik denk echter dat als je weet wat je aan elkaar hebt, die samenwerking er sowieso zal zijn. Bewijs daarvan zijn de vele co-publicaties van de afdeling met anderen. Daarnaast wil ik benadrukken dat wij in zo'n samenwerking een wat andere rol spelen, en moeten spelen, dan de overige co-auteurs. Wij moeten zonder voorbehoud kunnen zeggen: deze beweringen kunnen niet gestaafd worden aan de data of zijn onvoldoende danwel foutief onderlegd. U kunt zich voorstellen dat, zeker voor een junior statisticus, zo'n boodschap best lastig kan zijn als de laatste auteur je baas zou zijn en als het gaat om een manuscript dat hoog ingezet wordt.

Daarom ben ík blij dat het beleid van het VUmc is om biostatistiek te centraliseren. Dit is essentieel om onze poortwachtersfunctie te bewaken, hoogstaande kwaliteit te garanderen en voldoende kritieke massa te creëren om ook op statistiekvlak zeer wervend te zijn.

[DE CONCURRERENDE STATISTICUS]

Statistici zijn niet de enigen die genomische data analyseren. Bioinformatici doen dat ook. Laat ik eerst voorop stellen dat er veel complementariteit is tussen de twee expertises. Ik werk dan ook met veel plezier samen met de bioinformatici binnen het VUmc, juist omdat we goed van elkaar weten wat we wél en wat we niet kunnen. Op het internationale onderzoeksvlak is er echter wel een behoorlijke overlap tussen de twee expertises, en dat kan lastig zijn voor statistici die werken aan genomische data. Wij zijn immers in de minderheid. Ik vrees dat wanneer statistici onvoldoende bereid zijn om in de analyses contextuele informatie mee te nemen, zij de slag om invloedrijke publikaties gaan verliezen. Want die andere expertise, de bioinformatica, is juist heel sterk als het gaat om de biologische context.

Ook ik geloof dat goede wiskundige modellen met mooie theoretische eigenschappen vaak te prefereren zijn boven puur algoritmische alternatieven, zeker als het gaat om het generaliseren van de resultaten. Maar laat de concurrentie, tussen aanhalingstekens, ons scherp houden. Wij zullen open moeten staan voor hun methoden. Bovendien moeten we wiskundige elegantie niet als argument gebruiken wanneer het bruikbaarheid van de methode aangaat. Want daarmee overtuigen we misschien onszelf, maar niet de rest.

De contextuele kennis, maar ook ons toevalsinzicht geeft statistici werkend in een ziekenhuis een voorsprong op een andere opdoemende concurrent: de zogenaamde 'big data' analist. Een recent Volkskrant artikel besteedde uitgebreid aandacht aan een 'big data' analist die zijn algoritmes wil toepassen op grote genomische data sets. Op zich niks mis mee, maar het artikel bevatte twee forse misvattingen.

Ten eerste: er werd de indruk gewekt dat er in de ziekenhuizen helemaal geen expertise op dit vlak is. U begrijpt: dit was behoorlijk tegen mijn zere been. Maar, om de hand in eigen boezem steken: het betekent waarschijnlijk ook dat wij als statistici beter aan onze PR moeten werken. Ten tweede: de 'big data' analist impliceerde dat het verzamelen van data van meer individuen mogelijk leidt tot dé oplossing als het gaat om het voorspellen van bijvoorbeeld de kans op dikke darm kanker. Als statisticus erken ik natuurlijk meteen dat meer data altijd helpt. Maar ik weet ook dat de biologische werkelijkheid

vaak erg ingewikkeld is. Dus voor elk model of algoritme geldt: er is een plafond is aan de voorspelkracht met de huidige soorten data. Laat ik er verder niet meer over zeggen dan: ze verdienen een kans en ik ben benieuwd.

[DE DOCERENDE STATISTICUS]

In de academie komt kennis opdoen, via onderzoek of anderzijds, met de verantwoordelijkheid deze kennis te delen. Ik heb het genoegen om op Masternivo les te geven in Statistiek voor Hoog-Dimensionale data, een klasse data waartoe ook genomische data behoren. Dat doe ik zowel hier aan de VU, binnen het wiskunde-onderwijs, als in Leiden, in de Statistical Science Master. Het is heerlijk om een vak te mogen geven dat overlapt met je eigen onderzoeksinteresses. Maar wat is Statistiek voor Hoog-Dimensionale data? Tja..., eigenlijk alle, zeg maar gewone, statistiek, en dan nog meer vanwege het dimensie-probleem en de complexiteit van dit soort data. Dus, hoe breed maak je het? Hoeveel vertel je over de data? Wat doe je met software? Goed, we hebben het één en ander geprobeerd met wisselend succes, en hinkten wellicht in het begin nog te veel op twee benen.

Uiteindelijk was het goed om te bedenken waar ik zelf als student het meest aan heb gehad. Dat was niet zozeer het gebruik van software, of de details van een specifiek soort data danwel wiskundig model. Allemaal hele nuttige zaken, maar vaak goed zelf aan te leren als de situatie, bijvoorbeeld een bepaald project, erom vraagt. Nee, het zijn juist de generieke concepten die je goed moet doorgronden. Want als je díe begrijpt, dan heb je vaak al een raamwerk voor de oplossing van een praktisch probleem. Voor de precieze invulling komt het aan op je creatief vermogen en je wiskundige bagage. Daarmee kun je concepten toepassen op je eigen probleem en bestaande methoden aanpassen wanneer nodig.

Wat zijn die generieke concepten wat betreft Statistiek voor Hoog-Dimensionale data? Ik noem er een paar in het Engels, want de meesten laten zich slecht vertalen. Denk aan begrippen als ‘sparsity’ en ‘over-fitting’, technieken als ‘regularisation’, ‘multiple testing’ en ‘shrinkage’. Maar ook: computationele methoden als ‘double cross-validation’ en ‘singular value decomposition’. Als een statistiek student dit allemaal doorgrondt na de cursus, dan ben ik heel tevreden.

In het VUmc geven we om de twee jaar, in samenwerking met de Pathologie, een cursus Statistiek voor Genoomanalyse. Deze is voornamelijk bedoeld voor promovendi en postdocs. Zo mogelijk nog uitdagender voor ons vanwege de beperkte wiskunde- en programmeerachtergrond van de meesten van deze onderzoekers. Ze compenseren echter veel met een forse dosis wilskracht en arbeidsethos. Een deel is na deze cursus in staat standaardanalyses in ons favoriete statistische pakket R te doen, en de rest heeft in ieder geval wat opgestoken over de principes. Wellicht het belangrijkste is dat ze beter kunnen inschatten wat wel en wat niet mag. En ook: wanneer zelf aan de slag te gaan en wanneer onze hulp in te roepen.

Één retorische vraag heb ik nog wel: moet de opleiding Moleculaire Biologie niet veel kwantitatiever worden gemaakt met een aantal moderne statistiekvakken en een programmeercursus? Het is toch vreemd dat in dit omics tijdperk de gemiddelde afgestudeerde moleculair bioloog zich nog steeds moet

zien te redden met Excel. Natuurlijk zijn er de bioinformatici en statistici om hen te helpen, maar het is voor de onderzoekers zelf veel leuker als ze meer met hun eigen data kunnen.

[DE DANKBARE STATISTICUS]

Ik eindig met het dankwoord. Ik dank de besturen van de Stichting VU-VUmc, van de Stichting het Vrije Universiteitsfonds, van de faculteit Exacte Wetenschappen, alsmede het college van bestuur van de VU en de raad van bestuur van het VUmc voor het in mij gestelde vertrouwen.

Ik dank ook de mensen die betrokken zijn geweest bij mijn benoeming, danwel op bestuurlijk danwel op adviserend niveau. In het bijzonder Mathisca de Gunst en Bernard Uitdehaag voor hun initiatief tot deze benoeming, en Hans Brug voor zijn initiatief tot uitbreiding van de leerstoel naar het VUmc. Bij dit geheel betrek ik ook de voorzitter van vandaag, Ronald Meester.

Mathisca en Aad van der Vaart ben ik erkentelijk voor de samenwerking en het opbouwen van de Statistiek voor Levenswetenschappen hier aan de VU. Een goed voorbeeld van hoe mathematische en toegepaste statistiek samen kunnen gaan. Laten we samen met andere collega's in Nederland het schot tussen die twee hokjes weghalen.

Ik dank Hans Berkhof, de kapitein op het VUmc biostatistiek schip: voor het mede- opbouwen van die mooie club mensen die daar nu zit, en zeker ook voor het delen van je brede kennis op het gebied van statistiek. We vinden elkaar in het motto: “Wiskunde is prachtig, maar de data moeten leidend zijn."

Ik ben erg blij met mijn statistiek-voor-genomics team, en betrek daarbij ook spelers uit het verleden. Zonder team speel je immers een verloren wedstrijd. Een bijzonder woord van dank voor twee constante factoren in dit team: Renée de Menezes en Wessel van Wieringen. Gelukkig zijn jullie allebei eigenwijs genoeg om je maar af en toe iets van mij aan te trekken: precies waar ik voor sta in een academische omgeving.

Tot de collega's en ex-collega's bij wiskunde, bij de VU en bij de TU Eindhoven: samen leggen we het zo belangrijke fundament dat de wiskunde is voor mij, maar ook voor de wetenschap als geheel. Want, om de discussie over de toepassingswaarde van wiskunde gelijk te smoren: het Platform Wiskunde Nederland rekende onlangs uit dat maar liefst een kwart van de werkgelegenheid in Nederland te danken is aan wiskunde. Goed, misschien een lichte overschatting, maar toch...

Ik dank de vele VUmc collega's van CCA-Vici, de Medische Oncologie, de hoofd-halsoncologie en in het bijzonder de Pathologie. Van die laatste afdeling hebben verschillende secties zoals de tumor profiling unit, de moleculaire pathologie en de microarray faciliteit veel geïnvesteerd in samenwerking met ons. Dank voor jullie data, hypotheses, en enthousiasme voor genomics. En excuses voor al die keren dat ik als zeurderige statisticus dat enthousiasme de kop in heb gedrukt.

Tot slot: wiskunde zit niet in mijn genen. Het is een statistisch en genetisch wonder dat jullie, pa en moe, als alfa en zorgtype niet één, maar twee keer een bèta-zoon maakten. Dank daarvoor en voor alle andere goede zaken die ik heb meegekregen van jullie. En ok, natuurlijk ook een paar minder goede...

Tot die andere bèta, mijn broer Bas: onze gespreksonderwerpen zijn zeer divers, maar laten zich samenvatten als www: Wieletjes, Willem II, Wetenschap. Altijd fijn om een 'brother in arms' te hebben.

Ik heb mijn genen ook weer met succes doorgegeven aan mijn twee Wieletjes: Stijn & Niels. Heerlijk dat jullie als ik thuis kom de formules lekker uit mijn hoofd schudden. En tot Dianne, mijn eigen, favoriete biologe: je houdt niet van genomics, maar wel van mij. En dat is meer dan voldoende.

Ik heb gezegd.