*KLAAS FABER**

Het biologisch paspoort: veelbelovende opsporingstechniek of juridisch wankel?

Op grond van de herziene Wereld Anti-Doping Code is het met ingang van 1 januari 2009 mogelijk om een sporter énkel op basis van verdachte waarden in zijn/haar biologisch paspoort te vervolgen. In de standaardbehandelingen van dit nieuwe, indirekte bewijsmiddel wordt vooral gewezen op de extra mogelijkheden die deze aanpak biedt, naast het conventionele testen dat gewoon in stand blijft. Hier wordt gedetailleerd ingegaan op problematische aspecten die deels fundamenteel van aard zijn en dientengevolge niet eenvoudig te repareren. Voorts wordt aan de hand van de de zaak van Claudia Pechstein geïllustreerd dat dit nieuwe bewijsmiddel, mits zonder (deugdelijk) steunbewijs ingezet, een grote mate van willekeur in de hand werkt, waartegen de verdachte sporter zich nauwelijks kan verweren.

1. Inleiding

1.1. Setting the stage

Het biologisch paspoort staat momenteel volop in de belangstelling als de nieuwste ontwikkeling in de strijd tegen dopinggebruik in de sport. Kenmerkend voor deze ontwikkeling is, dat zij neerkomt op een principieel andere benadering van het testen op doping. Waar men bij het conventionele testen tracht om een verboden stof of methode specifiek en overtuigend aan te tonen, registreert men bij het biologisch paspoort veranderingen in bloedwaarden die gevoelig zijn voor doping. Die indicaties voor doping zijn namelijk langer waar te nemen dan de verboden stof of methode zelf, zodat de pakkans effectief wordt vergroot. Kortom: geheel indirect in plaats van direct bewijs.

De meningen over de wenselijkheid en toepasbaarheid van dit indirecte bewijs lopen sterk uiteen, zoals reeds wordt gesuggereerd in de titel van dit artikel. Deze is overigens ontleend aan een eerdere versie van het stuk van Asha ten Broeke[1] waarin het officiële standpunt ('veelbelovend') evenwichtig wordt afgezet tegen twee min of meer afwijkende meningen ('bedenkingen'). Ten Broeke heeft uiteindelijk gekozen voor de meer prikkelende titel 'Paspoort op glad ijs?', wellicht om aan te haken bij de actualiteit, namelijk het slepende proces tegen vijfvoudig Olympisch kampioen langebaanschaatsen Claudia Pechstein.

Dit artikel bestaat in essentie uit twee grote onderdelen. In sectie 2 wordt gedetailleerd ingegaan op diverse problematische aspecten van dit indirecte bewijs. Noem het de theorie. De mate van detail en diepte van behandeling zal de gemiddelde jurist wellicht afschrikken. Het betreft immers veelal (deel)onderwerpen die hun oorsprong vinden in een discipline die reeds op de middelbare school gezien wordt als een struikelvak, namelijk statistiek. Echter, de thans heersende, duidelijk oppervlakkige standaardbehandelingen kunnen makkelijk leiden tot onverantwoorde conclusies.

Sectie 3 is geheel gewijd aan de zaak van Claudia Pechstein, de eerste sporter in de geschiedenis die geschorst is op basis van dit indirecte bewijs in plaats van een 'positieve' test. De lezer zal ongetwijfeld kennis genomen hebben van de recente vrijspraak van Lucia de Berk,[2] de Haagse verpleegster die jarenlang van moord op meerdere patiënten is verdacht op grond van indirect bewijs dat uiteindelijk toch geen stand kon houden. In sectie 3 wordt o.a. betoogd dat de gebrekkige bewijsvoering in de zaken van Lucia de Berk en Claudia Pechstein ten minste op één cruciaal punt grote gelijkenis vertoont (zie 3.5). Secties 2 en 3 zijn zodanig geschreven dat men eerst sectie 3 kan lezen om, indien men verdere verdieping wenst, terug te keren naar sectie 2. Helaas impliceert deze opzet enige overlap doch hopelijk beklijft hierdoor de wat 'zwaardere' kost in sectie 2, zonder dat de herhalingen te veel storen.

Verregaande claims, met name met betrekking tot publieke (des)informatie, worden onderbouwd met referenties naar wetenschappelijke artikelen. Een illustra-

* Dr. N.M. Faber is eigenaar van Chemometry Consultancy, een adviesbureau op het gebied van data-analyse. Daarnaast onderhoudt hij contacten met diverse universiteiten voor het verder ontwikkelen en toepassen van methoden voor onderzoek.

1 E. van Laar, 'Paspoort op glad ijs?', *C2W* 23 januari 2010.

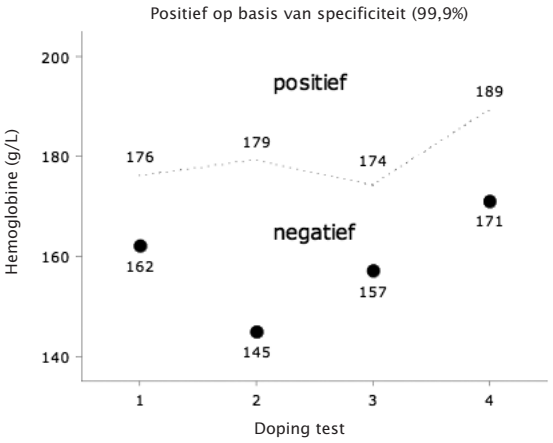
2 <www.luciadeb.nl/>; geconsulteerd op 19 april 2010.

tief voorbeeld dat wordt verspreid door de internationale wielerunie UCI, luidt als volgt:[3]

> 'The scientific assessment of a rider's profile applies similar principles to those used in forensic medical science to determine the likelihood of guilt.'

Deze publieke informatie zal blijken faliekant onjuist en daardoor extreem misleidend te zijn: bij de huidige stand van zaken is het een aantoonbaar gebrekkig bewijsmiddel.
Ten slotte zijn referenties naar diverse 'grijze' bronnen zoals krantenartikelen toegevoegd, alsmede links naar Wikipedia. Die bronnen zijn bedoeld om het totaalbeeld te completeren.

1.2. De steen des aanstoots: het biologisch paspoort
Onder het motto 'een plaatje zegt meer dan 1000 woorden' is in figuur 1 een grafiek uit een wetenschappelijk artikel over het biologisch paspoort[4] gedeeltelijk gereproduceerd. Men ziet hier vier (4) bloedwaarden ('●'), namelijk de concentratie van hemoglobine in gram per liter bloed, weergegeven in relatie met een stippellijn die de (statistisch-gefundeerde) beslisregel voorstelt.



Figuur 1

Bloeddoping leidt tot een toename van hemoglobine, waardoor op onnatuurlijke wijze een beter zuurstoftransport wordt bewerkstelligd. De beslisregel bepaalt dat men onder de stippellijn 'negatief' is voor bloeddoping, en daarboven 'positief'. De stippellijn is berekend voor een zogenoemde specificiteit van 99,9%.
Het begrip specificiteit wordt in sectie 2 uitgelegd. In de media vindt men hiervoor benamingen als 'zekerheidspercentage' en 'afkapgrens'. Vooral de eerste term is uiterst misleidend. In sectie 2 zal blijken dat de relatie met de kans op 'fout-positieven' geheel onduidelijk is, ook al suggereert die term 1 op 1000. De stippellijn in figuur 1 kan men derhalve opvatten als een hoogtelijn op een landkaart, waarbij niet duidelijk is welke hoogtes nu eigenlijk verdeeld worden. Sterker nog: voor verschillende gebergtes (sportdisciplines) zal die onbekende relatie verschillend zijn. Anders gezegd: de kans op schuld wordt in figuur 1 overschat en de mate van overschatting is per sportdiscipline verschillend (zie verder 2.8.1).
Dit bij voorbaat ongedefinieerd overschatten van de kans op schuld, oftewel de 'likelihood of guilt' in bovenstaande informatie van de UCI, is uiteraard een zorgwekkende constatering omdat het indirecte bewijs geheel op zichzelf staand tot een veroordeling kan leiden. Het is mijns inziens dan ook niet overdreven om te benadrukken dat de verdachte doorgaans als een zogenoemde 'cold hit suspect' beschouwd dient te worden. Intuïtief kan men dan reeds aanvoelen dat bijvoorbeeld deugdelijk steunbewijs aangedragen moet worden, een vereiste dat nu geheel niet wordt onderkend.

1.3. De ad-hocbeslisprocedure
In standaardbehandelingen wordt steevast benadrukt dat de uiteindelijke conclusie 'doping' niet enkel op statistiek is gebaseerd. Integendeel, op basis van het overschrijden van de statistisch-gefundeerde beslisregel vindt een 'initial review' plaats. Dit is mensenwerk. De uiteindelijke conclusie 'doping' is eveneens het resultaat van mensenwerk. Die laatste fase van het onderzoek wordt op p. 30 van de operating guidelines van het Wereld Anti-Doping Agentschap (WADA)[5] als volgt beschreven:

> 'Unanimous opinion of the panel that there is no known reasonable explanation for the blood profile information of this *Athlete* other than the use of a *Prohibited Substance* or *Prohibited Method*'.

De oplettende lezer herkent direct een klassieke drogredenering, namelijk het argument van de onwetendheid (argumentum ad ignorantiam),[6] ook wel negatief bewijs genoemd. Het spreekt vanzelf dat een drogredenering niet gehonoreerd dient te worden met een reglementaire status. Verder moge het duidelijk zijn dat zo'n tweestapsprocedure sowieso leidt tot 'hineininter-

3 <www.uci.ch/Modules/ENews/ENewsDetails.asp?MenuId=MjI0NQ&id=NTQzOA&LangId=1>; geconsulteerd op 19 april 2010.
4 N. Robinson, P.-E. Sottas, P. Mangin & M. Saugy, 'Bayesian detection of abnormal haematological values to introduce a no-start rule for heterogeneous populations of athletes', *Haematologica* 2007-8, p. 1143-1144.
5 <www.wada-ama.org/Documents/Science_Medicine/Athlete_Biological_ Passport/WADA_AthletePassport_OperatingGuidelines_FINAL_EN.pdf>; geconsulteerd op 19 april 2010.
6 <http://en.wikipedia.org/wiki/Argument_from_ignorance>; geconsulteerd op 19 april 2010.

pretieren' en *tunnelvisie*. Getallen gaan onvermijdelijk een eigen leven leiden, los van hun werkelijke betekenis die immers onbekend is. In sectie 3 wordt aan de hand van de zaak Pechstein duidelijk hoe extreem lastig het is om tegen een dergelijke ad-hocbeslisprocedure verweer te voeren.

2. Theorie: een 'crash course' forensische statistiek, rechtspsychologie en logica

2.1. Setting the stage

Een weinig bekend gezegde luidt: goede theorie is altijd praktisch. Een ander, eveneens weinig bekend gezegde luidt: goed onderzoek is duur, slecht onderzoek is onbetaalbaar. Om dergelijk krasse gezegdes op waarde te kunnen schatten, moet men helaas de nodige theorie tot zich nemen. De volgende uiteenzettingen beginnen waar standaardbehandelingen doorgaans eindigen. Ze vereisen dientengevolge aandachtig lezen en overeenkomstig geduld. Echter, op het eind van de rit volgt een aardige beloning: een beter begrip van dit aspect van dopingtesten dan aanwezig bij menig antidopingonderzoeker. Niet voor niets is op dit moment voor geen enkele test bekend wat de kans op 'fout-positieven' is.

2.2. De overtreding geconstateerd op grond van een 'abnormaal' testresultaat

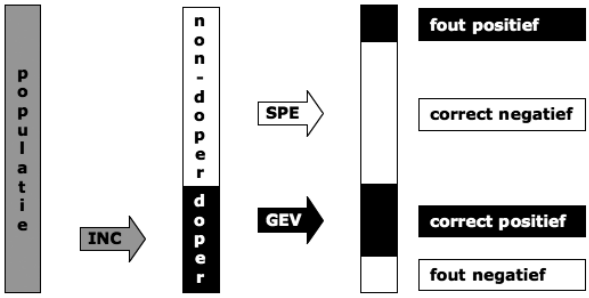
Laten we ons allereerst richten op een veelvoorkomende beslisregel in het conventionele testen. Voor het biologisch paspoort is de uitleg geheel analoog, zie figuur 1. Vervolging geschiedt op basis van een *'abnormaal' testresultaat*. Om te bepalen wat 'abnormaal' is, heeft men natuurlijk een beslisregel nodig. Voor lichaamseigen stoffen zoals nandrolon geldt een drempelwaarde. Een testresultaat onder die waarde wordt gezien als 'normaal' oftewel 'negatief', daarboven als 'abnormaal' oftewel 'positief'. Een 'normaal' resultaat is uiteraard in orde, 'abnormaal' (te hoog) geldt als een overtreding van het Dopingreglement.

2.3. De kans op schuld structureel overschat door de misleidende fixatie op 'abnormale' resultaten

Vaak gaat men als volgt te werk om de kans op schuld in te schatten. Er wordt een grote groep sporters bestudeerd, waarvan bekend is dat ze geen doping (nandrolon) gebruiken. De beslisregel ('normaal' ↔ 'abnormaal') wordt zodanig ingesteld dat slechts een kleine fractie als 'abnormaal' wordt beoordeeld – zeg 0,1%. Dat komt neer op een zogenoemde specificiteit van 99,9%, namelijk de kans op een 'correct-negatief' resultaat voor een 'schone' sporter. Men meent vervolgens 99,9% zeker te zijn van dopinggebruik indien er een 'positief' resultaat wordt gevonden, dus 0,1% kans op een 'fout-positief' voor een aangeklaagde sporter.
Dat hier een kolossale denkfout wordt gemaakt, blijkt reeds door die redenering op bovenstaande controlegroep toe te passen. Er is géén doping gebruikt en derhalve is men bij iedere 'positieve' uitslag nog steeds zeker van onschuld: elke 'positieve' uitslag is een 'fout-positief'.
Samenvattend: de kans op een 'fout-positief' van 0,1% geldt voor álle sporters vóórdat er getest wordt, maar níet voor de individuele sporter die er uitgepikt wordt als gevolg van een 'positief' resultaat. Die kans van 0,1% is derhalve *geheel niet relevant* in een concrete dopingzaak, want je wilt juist voor de aangeklaagde sporter de kans op schuld weten, en niet voor álle sporters die aan een test worden onderworpen.

2.4. De juiste berekening van de kans op schuld

De juiste berekening van de kans op schuld verloopt als volgt via de kans op een 'fout-positief' (samen opgeteld 100%). Voor het berekenen van de kans op 'fout-positieven' moet men drie kansen kennen,[7] zie figuur 2. De grootte van de afzonderlijke blokken is vooral illustratief bedoeld, dus zeker niet correct naar schaal. Met name de kans op 'fout-positieven' is in het algemeen zo klein dat dit blokje op correcte schaal slechts een ragfijn lijntje zou zijn.



Figuur 2

Die drie kansen zijn als het ware nodig om de populatie van atleten volledig uit te kunnen splitsen:
1. incidentie (INC): nodig om de populatie te splitsen in dopers en non-dopers;
2. specificiteit (SPE): nodig om de non-dopers te splitsen in 'correct-negatief' en 'fout-positief'; en
3. gevoeligheid (GEV): nodig om de dopers te splitsen in 'correct-positief' en 'fout-negatief'.

Hopelijk verduidelijkt figuur 2 de onvolmaakte, doch wellicht aanvaardbare opsplitsing van de populatie door middel van de test. Nog even ter herinnering: een 'positief' testresultaat leidt tot de conclusie 'doping'. In het toepassingsstadium kan natuurlijk geen onderscheid gemaakt worden tussen 'correct-positief' en 'fout-positief'. Het is derhalve zaak om vóóraf, dus

7 D.A. Berry & L. Chastain, 'Inferences about testosterone abuse among athletes', *Chance* 2004-2, p. 8-11.

vóór de introductie van een test, de inschatting te maken dat de fractie 'fout-positieven' aanvaardbaar is. Dit vereist objectieve en betrouwbare schattingen van alle drie kansen in figuur 2.
Tot slot: de gemiddelde jurist deelt zijn/haar afkeer voor statistiek met de meeste medici. Echter, medici krijgen deze fundamentele stof meer dan grondig onderwezen in verband met screening voor bepaalde aandoeningen. Bijgevolg zal een medicus de kans op een aandoening niet zomaar overschatten, zoals hier behandeld is voor de kans op schuld.

2.5. Rekenvoorbeeld

Dan nu een rekenvoorbeeld. Ga uit van een populatie van 10 000 atleten en:

1. incidentie = 10%. Er zijn derhalve 1000 dopers en 9000 non-dopers.
2. specificiteit = 99,9%. Dit geeft 0,1% kans op een 'positieve' test voor een non-doper.
3. gevoeligheid = 5%. Dit geeft 5% kans op een 'positieve' test voor een doper.

De specificiteit is dan heel behoorlijk. De gevoeligheid (= pakkans) is matig, want veel kleiner dan de optimale waarde van 100%. Echter, bij een grotere pakkans zal het snel niet meer gebruikt worden. Men loopt immers vroeg of laat onvermijdelijk tegen de lamp.
Het resultaat is: 0,1% x 9000 = 9 'fout-positieven' en 5% x 1000 = 50 'correct-positieven'. Totaal is 9 + 50 = 59 'positieven'. De kans op 'fout-positief' voor een aangeklaagde non-doper is derhalve 9/59 = 15% (tegen 0,1% voor álle 9000 non-dopers), terwijl de kans op schuld ('correct-positief') volgt als 50/59 = 85%. Is dat wel voldoende voor een veroordeling?

2.6. Wat is goed genoeg voor veroordeling in strafzaken?

In het strafrecht zijn 'verbale' schalen in omloop. Een bekende is van Hummel die nota bene in de context van het biologisch paspoort wordt behandeld:[8]

- Praktisch bewezen: 99,80%-99,90%
- Extreem waarschijnlijk: 99,10%-99,79%
- Zeer waarschijnlijk: 95,00%-99,09%
- Waarschijnlijk: 90,00%-94,99%
- Onbeslist: 80,00%-89,99%
- Niet nuttig: minder dan 80,00%

Het huidige resultaat (85%) is dus 'onbeslist' in een strafzaak. Voor het verhogen van dit percentage ligt het voor de hand om aan de gevoeligheid van de test te werken. Het is eenvoudig te controleren dat met een gevoeligheid van 10% (in plaats van 5%) het resultaat 92% zou zijn: 'waarschijnlijk'. Dat is nog steeds een stuk minder zwart-wit dan wat het laboratorium op dit moment rapporteert, namelijk 'praktisch bewezen'. De fractie 'fout-positieven' blijkt immers nog steeds 8%/0,1% = 80 maal groter te zijn dan hetgeen het laboratorium opgeeft.

2.7. Wat is goed genoeg voor veroordeling in dopingzaken?

Deze auteur is in een krantenartikel met de veelzeggende titel 'Dopingexpert Faber gelooft in toeval en eigen theorieën' als volgt gecompromitteerd:[9]

> 'Voor de arbiters hoeft het bewijs niet "boven iedere twijfel verheven" te zijn, zoals in strafzaken. Dopingovertredingen moeten aannemelijk worden gemaakt ("comfortable satisfaction"), dat is voldoende. Aannemelijk – het is een maatstaf die het wantrouwen van een complotdenker als Faber voedt.'

Laten we kijken in hoeverre die bewering een grond vindt in de huidige anti-dopingpraktijk zélf. De bronnen omtrent het biologisch paspoort reppen van een kans op een 'fout-positief' van 1 op 1000,[10] hetgeen neerkomt op 'praktisch bewezen'. Harm Kuipers, mede verantwoordelijk voor het biologisch paspoort bij de internationale schaatsbond ISU, spreekt eveneens van een 'afkapgrens' van 1 op 1000.[11] N.B. Deze waarde is een factor 10 hoger (!) dan degene die beschouwd wordt door Sonksen:[12]

> 'The sports authorities have never published what they consider an "acceptable risk" (for a false +ve) but a workshop in our GH-2000 project that included a senior IOC lawyer settled on a risk of 1:10,000 as being "acceptable" and this figure has subsequently been used in many discussions and publications.'

Schrijver dezes kent geen betrouwbare schattingen over het aantal 'fout-positieven' dat in de praktijk 'gerealiseerd' wordt, en het ligt voor de hand dat ze eenvoudigweg niet bestaan. Een redelijke aanwijzing is echter te ontlenen aan een evaluatie van de anti-dopingonderzoekers Van Eenoo en Delbeke[13] die opmerken dat het aantal betwiste 'positieve' A-monsters ligt tussen 0,5% en 1% (langetermijngemiddelde: 1996-2007). Bedenkend dat circa 2% van de A-monsters een

8 P.-E. Sottas, N. Robinson, M. Saugy & O. Niggli, 'A forensic approach to the interpretation of blood doping markers', *Law, Probability and Risk* 2008-3, p. 191-210.

9 M. van Driel, 'Dopingexpert Faber gelooft in toeval en eigen theorieën', *de Volkskrant* 7 december 2009.

10 Zie noot 8; K. Sharpe, M.J. Ashenden & Y.O. Schumacher, 'A third generation approach to detect erythropoietin abuse in athletes', *Haematologica* 2006-3, p. 356-363.

11 T. Zonneveld, 'Mag ik uw paspoort aub?', *Dagblad De Pers* 3 juli 2009.

12 <www.bmj.com/cgi/eletters/337/jul04_1/a584#200589>; geconsulteerd op 19 april 2010.

'positief' testresultaat geeft, leidt door middel van eenvoudige vermenigvuldiging tot de kans op een betwist A-monster die ruwweg ligt tussen 1 in 10 000 en 2 in 10 000. Deze range zou een redelijke bovengrens moeten bieden voor het werkelijk 'gerealiseerde' aantal 'fout-positieven'. Immers, indien dit laatste aantal (beduidend) hoger zou liggen, dan zouden er (aanmerkelijk) meer sporters gaan klagen. De simpele conclusie luidt: er zijn zeer plausibele aanwijzingen dat men tot nu toe in dopingzaken ruim de maatstaf haalt die in strafzaken vereist wordt, namelijk 'praktisch bewezen' volgens de schaal van Hummel (zie 2.6).
Men kan ook langs geheel andere weg redeneren. Onder de paraplu van het WADA worden jaarlijks ruim 200 000 conventionele testen uitgevoerd door circa 35 geaccrediteerde laboratoria. Een 'fout-positieve' test kan tot verlies van accreditatie leiden. Ergo: reeds met een kans zo klein als 1 op 10 000 lopen anti-dopinglaboratoria een onaanvaardbaar bedrijfsrisico.

2.8. Onderschatte problemen met de forensische statistiek

2.8.1. Bepaling van incidentie, specificiteit en gevoeligheid mogelijk omstreden

Bij de bespreking van figuur 2 is benadrukt dat men vóóraf, dus vóór de introductie van een test, moet beschikken over objectieve en betrouwbare schattingen van alle drie kansen. De reden is simpel: de waarde die men aan die kansen toekent, bepaalt uiteindelijk de berekende kans op schuld. Door subjectieve waarden in te vullen, gaat men eigenlijk min of meer op de stoel van de rechter zitten en dat is natuurlijk niet de bedoeling. Dit geldt met name voor de incidentie. Binnen het (Nederlandse) strafrecht is dit bij uitstek een kans die door de rechter wordt ingeschat, niet door deskundigen namens verdediging of OM.[14]
Een voor de hand liggende doch naïeve manier van schatten die men vaak aantreft binnen doping, komt simpelweg neer op het turven van het aantal geconstateerde overtredingen. Illustratief is het volgende fragment:[15]

> 'Sinds 2000 testte de ISU zo'n 1650 schaatsers. Urinetests leidden in die periode tot drie schorsingen, herinnert Kuipers zich: 1 op de 550 schaatsers is dopingzondaar. "In het algemeen kan ik stellen dat doping bij schaatsen geen groot probleem is, wat omvang betreft", zegt Kuipers.'

Het ligt (vanwege 'fout-negatieve' testen) voor de hand dat dit een (flinke?) onderschatting oplevert, hetgeen uiteraard door de verdediging gebruikt kan worden om een te lage kans op schuld te suggereren. Immers, hoe minder men voor incidentie invult in de berekening, hoe groter de resulterende kans op een 'fout-positief' en hoe kleiner de resulterende kans op schuld (samen opgeteld 100%).
Een betere strategie lijkt het interviewen van sporters en andere direct betrokkenen. Aangezien het gaat om 'sociaal ongewenst' gedrag, zijn wellicht speciale methodes zoals de 'randomized response'-techniek nodig om eerlijke antwoorden te verkrijgen.[16] Het interviewen van direct betrokkenen stond centraal in een onderzoek dat lange tijd in diverse landen de agenda van de anti-dopingbeweging heeft bepaald.[17] Illustratief is het volgende fragment:

> 'In her evidence to the US Senate Judiciary Committee Hearing on Steroid Abuse in America, chaired in April 1989 by Senator Joseph Biden Jr., Pat Connolly, a coach of the US women's track and field team, estimated that of the fifty members of the team at the 1984 Olympics, "probably 15 of them had used steroids. Some of them were medallists". Asked by Senator Biden whether the number of athletes using steroids had increased by the time of the Seoul Olympics of 1988, Connolly replied "Oh, yes. Oh, yes, it went up a lot". She estimated that "At least 40 per cent of the women's team in Seoul had probably used steroids at some time in their preparation for the games."'

Geheel in lijn met het voorgaande, kennen Berry en Chastain[18] als volgt aan sporters een actieve rol toe, overeenkomstig met die van de (Nederlandse, actieve) rechter in strafzaken:

> 'Prevalence of disease is relatively easy to estimate, depending on the patient population. Prevalence of substance abuse is not. There is an inevitable subjective aspect of assigning a prior probability. A hearing board made up of an athlete's peers is especially appropriate for making such assignments.' (Met 'prevalence' en 'prior probability' wordt incidentie aangeduid.)

Bij de beschrijving van figuur 1 is gesteld dat de kans op schuld wordt overschat en dat de mate van over-

13 P. Van Eenoo & F.T. Delbeke, 'Response on "Regulations in the field of residue and doping analysis should ensure a well-defined risk of a false positive declaration" by N.M. Faber', *Accreditation and Quality Assurance* 2009-4, p. 219-221.
14 M. Sjerps, 'Forensische statistiek', *Nieuw Archief voor Wiskunde* 2004-3, p. 106-111.
15 A. ten Broeke, 'De uitschieter van Pechstein', *Spits* 15 januari 2010.
16 W. Pitsch, '"The science of doping" revisited: Fallacies of the current anti-doping regime', *European Journal of Sport Science* 2009-2, p. 87-95.
17 C. Dubin, *Commission of enquiry into the use of drugs and banned practices intended to increase athletic performance*, Ottowa: Canadian Government Publishing Centre 1990.
18 Zie noot 7.

schatting ook nog eens per sportdiscipline varieert. Dat blijkt hier eenvoudig te volgen uit het gegeven dat de incidentie per sportdiscipline zal verschillen, mogelijk ook nog in de tijd (met name als gevolg van succesvol testen), zie getuigenis van Pat Connolly hierboven.
Aangaande het schatten van specificiteit en gevoeligheid zal ik betrekkelijk kort zijn. Er is traditioneel overwegend aandacht voor het schatten van specificiteit, wellicht omdat de benodigde onderzoeken verreweg het eenvoudigst zijn. Men hoeft immers enkel te turven hoeveel 'schone' proefpersonen 'positief' testen. Daarentegen moet men voor het schatten van gevoeligheid aan proefpersonen doping toedienen en dit stuit op ethische bezwaren. Verder is het zo dat topsporters een bepaalde genetische aanleg hebben waardoor zij verschillen van 'gewone' mensen. De voorkeur is derhalve om topsporters voor dergelijke onderzoeken te rekruteren. Dat impliceert echter nog een ander bezwaar: op het moment dat topsporters in het kader van een onderzoek doping gebruiken, overtreden ze het Dopingreglement.[19]
Als gevolg van onrealistisch opgezette studies kunnen met name schattingen van gevoeligheid (pakkans) omstreden zijn (onrealistisch hoog?). Met al deze complicaties dient men voor het schatten van de kans op schuld terdege rekening te houden.

2.8.2. Benodigd aantal metingen absurd laag ingeschat

Op p. 11 van de operating guidelines van het WADA[20] wordt gerept van een aanvaardbaar minimum van drie (3) punten. Hierover zal ik zeer kort zijn: iedere goede statisticus zal beamen dat dit aantal absurd laag is, zelfs voor het uitrekenen van zoiets simpels als een gemiddelde. Er zijn niet voor niets stelregels als 'twenty is plenty'. Uit een dergelijke onnadenkendheid valt af te leiden dat deze guidelines niet door een goede statisticus zijn opgesteld en/of gecontroleerd.

2.8.3. Vaker dan eenmalig testen niet verdisconteerd

Bij de berekening van de kans op schuld, zoals die hierboven in detail is uiteengezet, gaat men ervan uit dat de test slechts eenmalig wordt toegepast. Bij iedere test neemt echter de overallkans op een foutieve beslissing toe. Veronderstel dat eenmaal testen in 99% van de gevallen goed gaat. Dan gaat het na twee keer nog maar in 99% x 99% = circa 98% van de gevallen goed. Dit (bij statistici goed bekende) cumulatieve effect blijkt niet te zijn verdisconteerd in de methodologie achter het biologisch paspoort.[21] Het spreekt vanzelf dat de kans op schuld hierdoor aanzienlijk wordt overschat.

2.8.4. Overgesimplificeerde aanname omtrent de kansverdeling van de metingen

Een illustratief fragment uit een populair-wetenschappelijke verhandeling volgt:[22]

> 'Gill signaleert een nog fundamenteler probleem. Volgens hem zijn de berekeningen allemaal gebaseerd op de zogeheten Gauss-verdeling. Dit veronderstelt dat allerlei meetwaarden die een natuurlijke variatie vertonen, dat op een standaardmanier doen: bijna alle waarden in een kluitje rond het gemiddelde, terwijl uitschieters heel snel zeldzamer worden naarmate ze verder afwijken. Volgens Gill is dit niet meer dan een 19e-eeuws dogma, waarvan bewezen is dat het met name in medische kwesties vaak niet klopt. Het probleem is dat de juiste kansverdeling in veel gevallen niet goed experimenteel is bepaald. Wel is volgens Gill duidelijk dat natuurlijke uitschieters veel minder zeldzaam zijn dan volgens de Gauss-verdeling, zodat de kansen op "vals-positieven" nu schromelijk worden onderschat.'

Gill is de Leidse hoogleraar statistiek die een belangrijke rol heeft gespeeld in de heropening van de zaak van Lucia de Berk.

2.9. Onderschatte problemen met de rechtspsychologie en logica

Eigenlijk zijn de concrete (reken)problemen met statistiek een manifestatie van dieperliggende (psycho)logische problemen. Thompson en Schumann[23] komt de eer toe als eerste hierop gewezen te hebben, terwijl dichter bij huis Willem Wagenaar pionierswerk heeft verricht.[24] Die (psycho)logische problemen uiten zich met name ook in de structuur van uitspraken die door een (forensisch) deskundige kunnen worden gedaan.[25] Een illustratief voorbeeld is het fragment uit 1.1 waarin de UCI stelt dat het huidige biologische paspoort toelaat een uitspraak te doen over de kans op schuld.

19 B. Brouwer, H.F.M. Lodewijkx & H. Kuipers, 'Dopingbekentenis langs de wetenschappelijke meetlat', *Sportpsychologie Bulletin* 2009, nr. 20, p. 24-37.
20 Zie noot 5.
21 N.M. Faber & B.G.M. Vandeginste, 'Flawed science "legalized" in the fight against doping: the example of the biological passport', *Accreditation and Quality Assurance* 2010-6, p. 373-374.
22 A. Jaspers, 'Bloedschande. Dopingpaspoort als fabeltje', *Natuurwetenschap & Techniek* 2010-2, p. 34-35.
23 W.C. Thompson & E.L. Schumann, 'Interpretation of statistical evidence in criminal trials. The prosecutor's fallacy and the defense attorney's fallacy', *Law and Human Behavior* 1987-3, p. 167-187.
24 W.A. Wagenaar, 'The proper seat. A Bayesian discussion of the position of expert witnesses', *Law and Human Behavior* 1988-4, p. 499-510.
25 I.W. Evett, G. Jackson, J.A. Lambert & S. McCrossan, 'The impact of the principles of evidence interpretation on the structure and content of statements', *Science & Justice* 2000-4, p. 233-239.

Dat is logischerwijs uitgesloten: aangezien het (enige) bewijs geheel is gebaseerd op specificiteit, kan men enkel een uitspraak doen over de kans op een 'fout-positief' voor een 'schone' sporter. Díe uitspraak wordt echter niet gevraagd en dan blijkt de verleiding zeer groot te zijn om toch de gewenste uitspraak te doen, die echter niet door het (enige) bewijsmiddel wordt gedragen. Dit bekende probleem is verder door Faber en Sjerps uitgewerkt in het kader van het biologisch paspoort.[26] Last but not least is daar nog het argument van de onwetendheid, de drogredenering die nodig blijkt te zijn om het bewijs 'rond te krijgen' (zie 1.3).

2.10. Onderschatte problemen met de praktijk

Zoals eigenlijk te verwachten is bij iedere nieuwe methode, brengt ook de introductie van het biologische paspoort nieuwe praktische problemen met zich. Men dient hier te bedenken dat de methodologie vrijwel ongewijzigd is overgenomen uit toepassingsgebieden waar met name fraude niet voor de hand ligt, zoals medische screening.[27]

Dit is een relevante constatering want het is immers te voorzien dat dopingzondaars zullen trachten de bloedwaarden door aangepaste dopingregimes te manipuleren, zodat men binnen de toegestane bandbreedte blijft. Dat zou een effectieve manier zijn van 'onder de radar te gaan'. Heeft men namelijk geen opvallende bloedwaarden, dan komt men ook niet in aanmerking voor extra conventionele testen. Er zijn concrete aanwijzingen dat dit 'spel' reeds gaande is.[28]

Een ander curieus doch eveneens weinig verrassend euvel doet zich voor met zevenvoudig Tour de France-winnaar Lance Armstrong:[29]

> 'Volgens de Deense bloedspecialist zijn de bloedwaarden van de Amerikaan "ongeloofwaardig". (…) Het aantal rode bloedlichaampjes, de hematocrietwaarde én de hemoglobinewaarden bleven ook in de slotweek van de Tour hoog terwijl zijn hematocrietwaarde tussen 11 en 14 juli van 40,7 naar 43,1 steeg. Normaal gesproken dalen deze waarden tijdens een zware wedstrijd als de Tour de France. "Dat was bij Armstrong niet het geval. Dat kan dus duiden op het gebruik van bloeddoping", vertelde Mørkebjerg.'

De indirecte getallen spreken simpelweg niet voor zich. Het zal geen verbazing wekken dat hier een interessante markt ligt voor laboratoria die hun diensten kunnen aanbieden om dopingtests te omzeilen.[30]

2.11. Reacties op externe kritiek

De bekende biostatisticus Don Berry heeft reeds in een algemene context gewezen op het gevaar van het werken met enkel specificiteit, wellicht het belangrijkste verborgen gebrek van het biologisch paspoort.[31] In een response ontkent Sottas, de ontwikkelaar van het biologisch paspoort voor het WADA, categorisch dat met de verkeerde kans gewerkt wordt.[32] Dat blijkt dus geheel bezijden de waarheid te zijn, zoals men eenvoudig kan verifiëren in figuur 1 (ontleend aan een ander artikel waarvan Sottas medeauteur is).

De voor de hand liggende mogelijkheid van manipulatie door dopingzondaars wordt evenmin serieus genomen. Op de opmerking 'Wetenschappers als de Nederlander Klaas Faber denken dat het paspoort gemakkelijk te omzeilen is door bloedwaarden te manipuleren.' antwoordde Pat McQuaid, voorzitter van de UCI:[33]

> 'Wij hebben acht wetenschappers die aan het paspoort werken, en zij denken van niet. Mijn idee is dat het paspoort solide blijkt te zijn als het tot een rechtszaak komt.'

De lezer oordele zelf na het lezen van dit artikel.

2.12. Afsluitende opmerkingen

Diverse theoretische problemen met het nieuwe indirecte bewijs zijn hier de revue gepasseerd. De meerderheid van deze problemen ontbreekt in de inmiddels zeer talrijke standaardbehandelingen, is niettemin deels fundamenteel van aard (derhalve niet eenvoudig te repareren), en bovendien zeker niet zonder gevolgen voor de bewijsvoering. Hierdoor wordt met name de kans op schuld systematisch te hoog ingeschat. Praktisch gezien biedt dit indirecte bewijs overigens nieuwe kansen voor (met name bemiddelde) dopingzondaars om aan de aandacht van hun belagers te ontsnappen.

26 K. Faber & M. Sjerps, 'Anti-doping researchers should conform to certain statistical standards from forensic science', *Science and Justice* 2009-3, p. 214-215.

27 M.W. McIntosh & N. Urban, 'A parametric empirical Bayes method for cancer screening using longitudinal observations of a biomarker', *Biostatistics* 2003-1, p. 27-40.

28 M. van Driel & M. Misérus, 'Dus bloedwaarde zegt ook niet alles', *de Volkskrant* 27 mei 2009.

29 <www.wielerupdate.nl/wielernieuws/12640/verdachte-bloedwaarden-lance-armstrong/>; geconsulteerd op 19 april 2010.

30 <www.cyclingarchives.com/txtzfiche.php?berid=692>; geconsulteerd op 19 april 2010.

31 D.A. Berry, 'The science of doping', *Nature* 2008-454, p. 692-693.

32 P.-E. Sottas, C. Saudan & M. Saugy, 'Doping: a paradigm shift has taken place in testing', *Nature* 2008, nr. 455, p. 166.

33 D. Elshout, 'Ik houd niet van dit onderwerp', *Dagblad De Pers* 25 november 2008.

3. Praktijk: het experiment Claudia Pechstein

3.1. Setting the stage

Claudia Pechstein is de eerste sporter in de geschiedenis die geschorst is op basis van 'abnormale' bloedwaarden in plaats van een 'positieve' test. Inmiddels is een klein boek over deze (non)dopingzaak(?) te schrijven. Ik zal me in deze schetsmatige uiteenzetting hoofdzakelijk beperken tot de betreurenswaardige rol van statistiek in deze zaak én de uiterst dubieuze rol van een kroongetuige. Daarnaast hecht ik eraan te benadrukken dat Pechstein is aangeklaagd door de internationale schaatsbond ISU. De ISU volgt hier namelijk níet de operating guidelines van het WADA. Het is eenvoudig te verifiëren (zie 3.4) dat er geen zaak geweest zou zijn indien men deze guidelines wél (terstond) had toegepast.

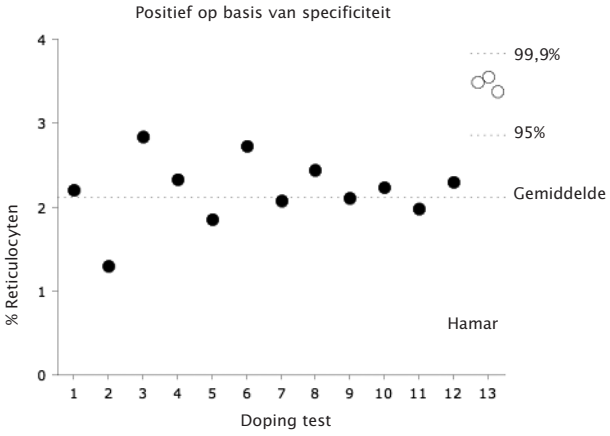
In sectie 2 is in detail uiteengezet dat de guidelines zélf cruciale gebreken vertonen. De verwarring lijkt dan ook compleet. Samenvattend kan men echter eenvoudig stellen dat Pechstein niet vervolgd zou zijn indien men de foutieve statistiek uit sectie 2 'correct' had toegepast. 'Doing the wrong thing right' was, zuiver pragmatisch beschouwd, goed genoeg geweest.

N.B. De verdediging heeft de toepasbaarheid van deze guidelines in haar verweer aangevoerd, maar die aanspraak op 'best practice' mocht niet baten daar de guidelines nog de status van draft hadden, zie punt 118 van de uitspraak van het Court of Arbitration for Sport (CAS).[34] Vanuit wetenschappelijk oogpunt is dat standpunt absoluut onacceptabel daar de guidelines zijn gebaseerd op onderzoek dat reeds lang was afgesloten. Deze guidelines zijn overigens één week na de uitspraak van het CAS zonder noemenswaardige wijzigingen goedgekeurd, namelijk op 2 december 2009, ofschoon het WADA vervolging op basis van indirect bewijs per 1 januari 2009 mogelijk heeft gemaakt. Het zijn met name de 'niet-technische' aspecten aan deze zaak die onvermijdelijk tot gevolg hebben dat hier de strikt wetenschappelijke beschrijving uit sectie 2 plaats moet maken voor een enigszins anekdotisch getint verhaal. Het zijn tevens de 'niet-technische' aspecten die maken dat het recente boek van Wagenaar, Israëls en van Koppen[35] voor menig stakeholder verplichte kost zou moeten zijn. Het volgende fragment is mijns inziens zeker van toepassing:

> 'Het gaat vooral om een beperkte categorie uitdagende zaken (…). In die zaken kunnen wij opsporen hoe er in juridische redeneringen met feiten wordt omgesprongen. En dan sporen wij geen incidenten op, geen bizarre gebeurtenissen, maar indicatoren van continu aanwezige structurele tekortkomingen.'

3.2. De verdachte bloedwaarden

De veroordeling van Pechstein is gebaseerd op een (eenmalige) verdachte schommeling van het percentage jonge rode bloedlichaampjes oftewel reticulocyten. Bloeddoping stimuleert de productie van reticulocyten, waarna hun rijping uiteindelijk leidt tot een verhoging van het gehalte van hemoglobine en bijbehorend zuurstoftransport. Figuur 3 geeft de relevante waarden vanaf 24 november 2007. Sommige waarden zijn gemiddeld omdat deze te kort na elkaar gemeten zijn, en derhalve niet onafhankelijk beschouwd dienen te worden.



Figuur 3

Het is evident dat de waarden die op 6 en 7 februari 2009 te Hamar gemeten zijn ('o': 1, respectievelijk 2 punten) 'abnormaal' hoog zijn ten opzichte van de rest ('●'). Maar hoe 'abnormaal' zijn ze?

Laten we allereerst figuur 3 kritisch, puntsgewijs nader beschouwen:

1. Om te beginnen een korte toelichting op het begrip specificiteit. Specificiteit is de kans op een 'correct-negatieve' uitslag voor een 'schone' sporter. In figuur 3 worden de waarden van Hamar namelijk vergeleken met de overige waarden die als natuurlijk worden gezien. In 2.3 wordt uitgelegd dat specificiteit door anti-dopingonderzoekers traditioneel wordt verward met de kans op schuld. Dat is niet zonder gevolgen; het rekenvoorbeeld in 2.5 laat zien dat de kans op schuld aldus wordt overschat.
2. Het gemiddelde van 2,1% is an sich niet 'normaal'. Verreweg de meeste mensen zitten op circa 1% gemiddeld. Bij Pechstein blijkt echter sprake te zijn van een aandoening waardoor zij gemiddeld ongeveer 2x zo hoog zit. Dit verhoogde niveau is verder niet relevant voor de zaak, want het is door het CAS geaccepteerd, zie met name punt 179 van de CAS-

34 <www.tas-cas.org/d2wfiles/document/3802/5048/0/FINAL%20AWARD% 20PECHSTEIN.pdf>; geconsulteerd op 19 april 2010.

35 W.A. Wagenaar, H. Israëls & P.J. van Koppen, *De slapende rechter*, Amsterdam: Bert Bakker 2009.

uitspraak. Pechstein mag dientengevolge alle titels en prijzen behouden die vóór Hamar zijn behaald.

3. Uit punt 183 van de CAS-uitspraak blijkt dat een ondeugdelijke berekening is gehanteerd voor specificiteit, namelijk het zogenoemde maximal critical difference.[36] Bovendien is die verkeerd berekende specificiteit met de verkeerde norm vergeleken, namelijk 95% 'zekerheid' terwijl (minimaal) 99,9% aangewezen is (zie 3.4).
4. Bij Pechstein was het gehalte hemoglobine niet verdacht, ook al is het verhogen van dit gehalte het uiteindelijke doel van bloeddoping. Deze parameter werd echter niet in de bewijsvoering meegenomen omdat ieder gehalte te manipuleren is door het bloed te verdunnen (hemodilutie). Bloedverdunning heeft daarentegen geen effect op het percentage reticulocyten omdat teller en noemer in gelijke mate worden beïnvloed. Hoe Pechstein deze manipulatie had moeten uitvoeren tijdens de talrijke bij haar uitgevoerde, onaangekondigde out-of-competition-controles, is echter geheel niet duidelijk gemaakt door de deskundigen van de ISU. Hier is dus zonder meer sprake van confirmatiebias,[37] meer in het bijzonder selectiebias. N.B. De operating guidelines van het WADA gaan uit van twee parameters, namelijk gehalte hemoglobine(!) en een samengestelde parameter die o.a. het percentage reticulocyten bevat (zie p. 28). Die samengestelde parameter is reeds in 2003 gepubliceerd.[38] Kortom: tunnelvisie in optima forma.
5. In Hamar zijn drie (3) metingen gedaan binnen twee dagen, wellicht uit het oogpunt van zorgvuldigheid. Die metingen zijn echter zeker niet onafhankelijk en wekken daardoor geheel onterecht de schijn van een vergrote zekerheid. Op p. 11 van de guidelines wordt overigens een tussenperiode van minimaal vijf dagen aanbevolen. In feite was in Hamar sprake van een denkfout die bekend staat als de Texaanse scherpschutter:[39] na de eerste meting werd een schietschijf rond deze meting geschilderd en nogmaals van korte afstand geschoten. Niet geheel toevallig raakten die laatste twee metingen doel. Deze denkfout is uiterst verleidelijk en komt daardoor in alle geledingen van de maatschappij voor, met name om áchteraf opzienbarende ontwikkelingen te ontdekken en daar vervolgens beleid op te richten.[40]

3.3. Alles verkeerd

Samenvattend: afgezien van de aantoonbaar subjectieve keuzen, is bij Pechstein de statistiek op werkelijk bizarre wijze misbruikt. Er is: (1) op definitieniveau gewerkt met de *verkeerde kans* (namelijk specificiteit) waarvoor (2) een getalletje ingevuld dat uit een *verkeerde berekening* (namelijk artikel van Banfi) komt rollen. Uiteindelijk is dit (3) totaal betekenisloze resultaat ook nog eens vergeleken met de *verkeerde norm* (namelijk 95% in plaats van 99,9%).
Kortom: hoe weet je nou of er echt sprake is van een verdachte uitschieter tijdens het toernooi in Hamar?

3.4. De verkeerde norm

Laten we allereerst naar die norm kijken. Daarvan is de functie uiteraard om het vertrouwen te kwantificeren dat je mag toekennen aan de conclusie 'doping'. In het kader van het biologisch paspoort wordt doorgaans gerept over een 'afkapgrens' van 1 op 1000 oftewel een 'zekerheidspercentage' van 99,9% (zie 2.7). Olivier de Hon, wetenschappelijk beleidsmedewerker van de Dopingautoriteit, stelt zelfs dat in deze zaak met een 'zekerheidspercentage' van 99,99% is gewerkt.[41] Dat laatste is dus aantoonbaar onjuist. Men kan zélf aan de hand van figuur 3 verifiëren dat het CAS-panel zich heeft laten leiden door een 'zekerheidspercentage' van 95%. Het is met name deze veel te slappe norm die grote aandacht heeft gekregen in de media.[42] Deze auteur is daar zelf geheel verantwoordelijk voor. De reden om juist díe fout te benadrukken is, dat de andere fouten veel moeilijker uit te leggen zijn (in met name interviews) én omdat het 'correct' gebruiken van de 'gangbare' norm (99,9%) reeds een beslissend verschil had gemaakt in deze zaak. Sterker nog: indien men in Hamar op 6 februari 2009 met de gebruikelijke norm had gewerkt, dan was er nooit een zaak geweest. Het is te verwachten dat de laatste twee fouten (verkeerde berekening en verkeerde norm) min of meer uitzonderlijk zijn in de zin dat ze niet herhaald zullen worden. Zó onnadenkend zal men toch niet zijn?!

3.5. De verkeerde kans: eerder Sally Clark en Lucia de Berk en dan nu Claudia Pechstein?

Laten we dan nu de aandacht richten op de eerste fout, het op *definitieniveau* werken met de verkeerde

36 G. Banfi, 'Reticulocytes in sports medicine', *Sports Medicine* 2008-3, p. 187-211.
37 <http://en.wikipedia.org/wiki/Confirmation_bias>; geconsulteerd op 19 april 2010.
38 C.J. Gore, R. Parisotto, M.J. Ashenden, J. Stray-Gundersen, K. Sharpe, W. Hopkins, K.R. Emslie, C. Howe, G.J. Trout, R. Kazlauskas & A.G. Hahn, 'Second-generation blood tests to detect erythropoetin abuse by athletes', *Haematologica* 2003-3, p. 333-344.
39 <http://en.wikipedia.org/wiki/Texas_sharpshooter_fallacy>; geconsulteerd op 19 april 2010.
40 M. Keulemans, 'Beheersbaar', *de Volkskrant* 9 januari 2010.
41 Zie noot 1.
42 M. Scholten & R. Schoof, 'Zaak Pechstein is gebaseerd op drijfzand', *NRC* 5 december 2009.

kans (namelijk specificiteit). Die tamelijk *abstracte* fout is inherent aan het biologisch paspoort – noem het een verborgen gebrek – en leidt tot de vergelijking van Claudia Pechstein met Sally Clark[43] en Lucia de Berk,[44] beiden triest voorbeeld van misbruik van statistiek in een strafzaak.[45] Sally Clark: tweemaal wiegendood, zó 'abnormaal' dat het geen toeval kon zijn, dus tweevoudige moord. Levenslang. Na drie jaar heropening en vrijspraak. Geen happy end:[46]

> 'Sally werd op 16 maart 2007 dood gevonden in haar huis in Hatfield Peverel. Oorspronkelijk werd gedacht dat ze aan een natuurlijk dood was gestorven. Later bleek dat ze was gestorven aan acute alcoholvergiftiging. De lijkschouwer vond geen bewijs voor zelfmoord. Ze liet haar man en haar derde zoon achter.'

Dichter bij huis speelt de zaak Lucia de Berk waarin eveneens de kans op schuld schromelijk is overschat door met de verkeerde kans te werken. Lucia de Berk is uiteindelijk op 14 april 2010, zeven jaar na haar veroordeling tot levenslang, vrijgesproken.[47]

3.6. De misleidende fixatie op 'abnormale' resultaten

Dan nu een belangrijke take home message: het op definitieniveau werken met de verkeerde kans hangt een-op-een samen met de thans allesbepalende doch misleidende fixatie op 'abnormaal'.

Een eenvoudig voorbeeld kan wellicht verduidelijken hoe soepel op de automatische piloot een onverantwoorde conclusie kan worden getrokken uit een 'abnormaal' resultaat. Stelt u zich een damesloterij voor met 1000 loten, waarbij slechts één prijs wordt uitgekeerd. Het langs eerlijke weg winnen van die loterij is redelijk 'abnormaal'. Met die stelling kunt u het onmogelijk oneens zijn: het is gewoon 'normaal' om te verliezen – de overgrote meerderheid doet het u dagelijks voor. Toch wordt op de winnares niet automatisch, zoals bij Pechstein is gedaan, de volgende redenering toegepast: de kans om op eerlijke wijze deze loterij te winnen is 0,1% (1 op 1000), dús heeft ze met een 'zekerheidspercentage' van 99,9% vals gespeeld.

Dat deze gedachtegang veel te kort door de bocht is, is reeds schematisch behandeld in 2.4. De allesbepalende fixatie op een 'abnormaal' testresultaat komt neer op het werken met de specificiteit van de test. Die specificiteit is hierboven aangeduid als 'zekerheidspercentage', terwijl Harm Kuipers, mede verantwoordelijk voor het biologisch paspoort bij de internationale schaatsbond ISU, verwijst naar de volstrekt analoge 'afkapgrens'. Laten we voor het gemak de veelbelovende term (sic!) 'zekerheidspercentage' aanhouden.

De in 2.4 gepresenteerde redenering is algemeen geldig, want gebaseerd op een elementaire stelling uit de wiskunde (Bayes). Het rekenvoorbeeld in 2.5 waarbij de fractie 'fout-positieven' 80 maal hoger uitvalt in 2.6 dan verwacht op basis van het 'zekerheidspercentage', zou derhalve ook op een toepassing van het biologisch paspoort kunnen slaan.

3.7. Prosecutor's fallacy

Het op deze klassieke manier overschatten van de kans op schuld staat binnen de rechtspsychologie bekend als prosecutor's fallacy: je krijgt er namelijk iemand makkelijk mee achter de tralies. Dit is gebeurd in de zaken van Sally Clark en Lucia de Berk. Zonder omhaal kan ik derhalve stellen dat het CAS van het vraagteken in de titel van 3.5 een uitroepteken heeft gemaakt.

Ik hecht er aan nogmaals te benadrukken dat dit overschatten van schuld in de context van dopingzaken reeds zeer expliciet is verwoord door de bekende biostatisticus Don Berry. Sottas, de onderzoeker die verantwoordelijk is voor de ontwikkeling van het biologisch paspoort voor het WADA, verstrekte daarop geen eerlijk antwoord, zie 2.11.

3.8. Het klassieke misbruik van statistiek aan de kaak gesteld

Vanwege een relevante wetenschappelijke publicatie,[48] waarin met name wordt voortgeborduurd op fundamenteel werk van vermaard rechtspsycholoog Willem Wagenaar, ben ik in juli 2009 door de verdediging van Pechstein gevraagd om een expert opinion te schrijven. Dat stuk moest zich vooral toespitsen op de statistische beoordeling van de bloedwaarden van Pechstein. Ik heb met name Sottas, één van de belangrijkste getuigen van de ISU, gevraagd waarom hij de gegevens op de klassieke manier mishandelde. Dat was voor mij, o.a. voortbouwend op de kritiek van Don Berry, makkelijk 'scoren'. Let trouwens eens op de titel van dat artikel ('Anti-doping researchers should conform to certain statistical standards from forensic science'): anti-dopingonderzoekers presenteren het biologisch paspoort graag als een forensische aanpak. Het tegendeel is eerder waar: bij de huidige stand van zaken is het een aantoonbaar gebrekkig bewijsmiddel.

43 <http://nl.wikipedia.org/wiki/Sally_Clark>; geconsulteerd op 19 april 2010.

44 Zie noot 2.

45 M. Buchanan, 'Conviction by numbers', *Nature* 2007, nr. 445, p. 254-255.

46 Zie noot 43.

47 Zie noot 2.

48 Zie noot 26.

3.9. Kroongetuige is zonder opgave van redenen teruggetrokken door de ISU

Zes weken na indienen van dat stuk kwam het CAS met haar uitspraak. Zoals bekend, handhaafde het CAS de schorsing. Dat mijn expert opinion niet geheel zonder effect is gebleven, blijkt echter uit punt 44 van de CAS-uitspraak:

> 'By communications faxed on 23 and 24 November 2009, the Athlete submitted an urgent application for the reopening of the hearing in order to have the opportunity to cross-examine Prof. Sottas, who had not attended the hearing of 22-23 October 2009. The reason for this application was that one of the Athlete's attorneys had apparently learned that Prof. Sottas had revised his previous opinion on the basis of the Appellants' evidence submitted on 14 October and, for that reason, the Respondent had not summoned him to the hearing. The Panel has taken into account the Athlete's application and has determined to dismiss it because, in reaching its decision, the Panel has not relied on the written expert opinion provided by Prof. Sottas.'

Sottas, die overigens helemaal geen professor is, had toch gewoon ter zitting kunnen uitleggen waarom hij de waarde van het *enige* bewijsmiddel overschat? Ik kan er op deze manier, zeker gezien de hieraan voorafgaande aantoonbare misleiding in wetenschappelijke artikelen (zie 2.11), enkel 'bedrog' van maken.

3.10. Reacties op externe kritiek

Harm Kuipers heeft een deel van de hierboven beschreven kritiek doorgespeeld naar de verantwoordelijke onderzoekers,[49] met als resultaat:

> 'Er is door statistici naar de bezwaren van Faber gekeken. Ze waren niet onder de indruk.'

Hierbij is het interessant te weten dat bijvoorbeeld Sottas, de onderzoeker die verantwoordelijk is voor de ontwikkeling van het biologisch paspoort, van huis uit geen statisticus is, maar bioloog.

3.11. Afsluitende opmerkingen

Claudia Pechstein is geschorst wegens 'abnormale' bloedwaarden die uitsluitend te verklaren zouden zijn door doping. Bij nader inzien zou een bizar misbruik van statistiek weleens met afstand het meest 'abnormale' aan deze zaak kunnen zijn. Ik ga er derhalve van uit dat het laatste woord in deze zaak voorlopig nog niet is gesproken. We zijn nu immers slechts getuige van de zichtbaar falende juridische toets terwijl de wetenschappelijke toets in feite nog grotendeels *in het openbaar* moet plaatsvinden.
Frappant aan deze zaak is dat enkel het hanteren van de in de huidige context voorgestelde, strakkere norm (99,9% in plaats van 95%) reeds een beslissend verschil had gemaakt.

4. Conclusie en aanbeveling

Het WADA spreekt van een 'betrouwbaar middel'.[50] De UCI claimt dat de kans op schuld wordt bepaald zoals gebruikelijk in 'forensic medical science' (zie 1.1). Een nadere beschouwing leert dat gebrekkige statistiek (zie sectie 2) wordt aangevuld met een bekende drogredenering (argument van de onwetendheid) om het bewijs 'rond te krijgen' (zie 1.3).
Volgens Harm Kuipers is statistiek de basis van deze methodologie. Aangezien een keten niet sterker kan zijn dan de zwakste schakel, moet de conclusie luiden dat dit 'betrouwbaar middel' niet voldoet voor vervolging. Ik zou dan ook graag de situatie van vóór 1 januari 2009 hersteld zien: gebruik het voor screening en ga aanvullend conventioneel testen indien er een verdenking rijst. Ik voeg aan de wetenschappelijke bezwaren enkele algemene bedenkingen toe om die aanbeveling te ondersteunen:

- De reacties op externe kritiek geven geen aanleiding om een wezenlijke verbetering van de methodologie te verwachten.
- Er is in dopingzaken géén adequate scheiding van de machten. Dit euvel is herkend in de zaak van Pechstein,[51] waar de internationale schaatsbond ISU diverse petten op heeft, maar het speelt zeker ook in het algemeen.[52]
- Sporters krijgen geen eerlijk proces in dopingzaken.[53] Dit probleem wordt versterkt door het (zonder steunbewijs) inzetten van indirect bewijs waartegen logischerwijs geen tegenbewijs mogelijk is.

Een voortzetting van de oude praktijk zou derhalve wel eens het meest wenselijke kunnen zijn, met als concreet voorbeeld:[54]

> 'Armstrong onderging dit jaar al meerdere dopingcontroles maar heeft nog geen enkele maal positief getest. Toeval of niet, ook vanmorgen onderging de Amerikaan een controle. Om tien voor zeven 's ochtends stonden de controleurs voor zijn deur in Austin, Texas.'

49 T. Zonneveld, 'Er dreigt een grote crisis', *Dagblad De Pers* 23 oktober 2009.
50 <www.wada-ama.org/rtecontent/document/code_v2009_En.pdf>; geconsulteerd op 19 april 2010.
51 <www.nzz.ch/nachrichten/sport/aktuell/keine_gewaltentrennung_im_fall_pechstein_1.5318078.html>; geconsulteerd op 19 april 2010.
52 <www.sportknowhowxl.nl/index.php?pageid=detail&catid=case-closed &cntid=3476>; geconsulteerd op 19 april 2010.
53 N.M. Faber, 'Pleidooi voor een eerlijk proces in dopingzaken', *Nederlands Juristenblad* 2009-44/45, p. 2880-2883.
54 Zie noot 29.