Dr. dr. R.W.M. Giard[*]

Overlijdensschade bij klinisch-wetenschappelijk onderzoek

Over waarheidsvinding in de woelige wateren van de wetenschap

1. De gevolgen van een proactieve persconferentie

Tijdens een indrukwekkende persconferentie maakten de coördinatoren van de PROPATRIA-studie op 23 januari 2008 wereldkundig dat tijdens dit medisch-wetenschappelijk onderzoek bij 296 patiënten met een ernstig verlopende alvleesklierontsteking er onverwacht meer mensen gestorven waren in de groep die experimenteel behandeld werd met een mengsel van bepaalde bacteriën (probiotica[1]) dan in de onbehandelde groep.[2] In deze studie stierven in totaal 33 patiënten: 24 in de onderzoeksgroep en 9 in de controlegroep, een oversterfte dus van 15 personen. De onderzoekers waren geschrokken van dit eindresultaat. In tegenstelling tot hun positieve verwachtingen had dit middel averechts gewerkt met overlijdensschade als gevolg en dat was voor hen een reden tot openbare reflectie.

Al direct na het naar buiten brengen van die uitkomsten, vier weken voor het verschijnen van de officiële wetenschappelijke publicatie in *The Lancet*[3], werden er in de media kritische vragen gesteld. Was dit onheil niet te voorkomen geweest? De Britse hoogleraar in de wiskundige statistiek van de Universiteit Leiden, Richard Gill, sprak in de tv-actualiteitenrubriek NOVA een vernietigend oordeel uit over de gang van zaken. De Inspectie voor de gezondheidszorg kondigde een onderzoek aan en enkele nabestaanden van overleden proefpersonen meldden dat zij een civiele procedure zouden gaan starten. Of je nu vanuit juridisch of niet-juridisch perspectief naar deze kwestie kijkt, er zal over dit – achteraf bezien – zo ongunstig verlopen onderzoek een normatief oordeel moeten worden uitgesproken met als centrale vraag: hebben de onderzoekers wel of niet verwijtbaar onzorgvuldig gehandeld?

Elders in dit nummer beschouwen Frijters en Van Houte deze trial en komen daarbij onomwonden tot een negatief oordeel over de manier waarop dit onderzoek werd uitgevoerd. Hun betoog steunt in belangrijke mate op een publicatie van de eerder genoemde statisticus Gill.[4] Wetenschapsbeoefening is gebaat met kritische confrontaties. Zachtzinnigheid is daarbij niet het primaire vereiste. Wijsheid wél, maar dan géén wijsheid achteraf. Als de vraag naar onzorgvuldigheid bij opzet en uitvoering van deze trial als oorzaak van overlijdensschade aan de rechter zou worden voorgelegd, hoe zou de ideale deskundige die deze rechter zou moeten adviseren de kwestie dan dienen te onderzoeken? In mijn artikel staat dit als thema centraal. Dat vraagstuk kan verder worden uitgewerkt in a) wat zijn bij de normatieve beoordeling van deze trial de kernvragen, en b) met welke methode(n) kunnen die beantwoord worden? En vanzelfsprekend: wat zijn dan de conclusies?

2. De kernvragen

We hebben te maken met een – schrikt u niet! – gerandomiseerde dubbelblinde placebogecontroleerde trial (hierna: RCT). Achter deze lange en ingewikkelde naam gaat een eenvoudig maar belangrijk principe schuil: door het lot worden twee groepen bepaald waarbij de ene behandeld wordt mét het middel van keuze en de ander geen werkelijke behandeling krijgt maar wel een vergelijkbare pil of drankje, de placebo. Hiermee wordt voldaan aan de in de klinische epidemiologie zo belangrijke *ceteris paribus doctrine*: de uitkomsten in beide groepen kunnen worden vergeleken omdat we aannemen dat ze maar in één opzicht van elkaar verschillen, namelijk of ze wel of niet de behandeling gekregen hebben. Zo kunnen valide conclusies worden getrokken over de effecten van de interventie. Voorwaarde daarvoor is dan wel dat de groepen voldoende groot zijn en dat er op het proces van randomisatie en deelname weinig of niets valt aan te merken. Noch de patiënt die deelneemt, noch de arts die de behandeling voorschrijft heeft er weet van of uiteindelijk het echte of het nepmiddel wordt toegepast: dat heet 'dubbelblind'. Aldus worden psychologische verwachtingseffecten tot een minimum beperkt.

Waar mogelijk wordt ernaar gestreefd om aan de hand van een RCT antwoord te krijgen op belangrijke diagnostische of therapeutische kwesties: een goed uitgevoerde gerandomiseerde trial geeft immers de meest betrouwbare antwoorden op medische vragen. Het alternatief is het doen van 'gewoon' observationeel onderzoek, dus géén experimenteel vergelijkend experiment. Er wordt een groep patiënten met een bepaalde aandoening behandeld en vervolgens wordt naar de effecten gekeken. De conclusies uit observationeel onderzoek zijn minder betrouwbaar.

Sinds invoering van de RCT als superieure mogelijkheid voor effecttoetsing van medische interventies is inmiddels vele malen gebleken dat veronderstelde heilzame effecten, geconstateerd in observationele studies, via RCT's werden ontkracht en zelfs schadelijk waren. Een

---

* Dr. dr. R.W.M. Giard is patholoog en klinisch-epidemioloog, Medisch Centrum Rijndmond-Zuid, jurist Rotterdam Institute of Private Law, Erasmus Universiteit Rotterdam en tevens redacteur van dit tijdschrift.
Met dank aan Diederik Aben en Ton Broeders voor hun commentaren op een eerdere versie van dit artikel.

1. Probiotica worden gedefinieerd als 'levende micro-organismen, die de gezondheid van de gastheer bevorderen indien toegevoegd in adequate hoeveelheden'.
2. *NRC Handelsblad* 24 januari 2008, 'Toen verschil bleek, werd het doodstil'.
3. *The Lancet* (371) 2008, p. 651-659; het artikel is ook in Nederlandse vertaling verschenen: M.G.H. Besselink et al., 'Probioticaprofylaxe bij voorspeld ernstige pancreatitis: een gerandomiseerde, dubbelblinde, placebogecontroleerde trial', *Nederlands Tijdschrift voor Geneeskunde* (152) 2008-12, p. 685-696.
4. R.D. Gill, 'Statistics, ethics, and probiotica', te downloaden via <http://arxiv.org/abs/0804.2522v1>.

bekend voorbeeld is het (langdurig) geven van hormonen aan vrouwen met overgangsklachten. Vele observationele studies toonden niet alleen afname van de hinderlijke klachten maar tegelijk werd een beschermende werking tegen hart- en vaatziekten waargenomen. RCT's maakten vervolgens het tegendeel duidelijk; hormoonbehandeling had geen beschermend effect op vaten en bleek zelfs schadelijk, want de kans op kanker nam toe.

Mag je dan altijd 'blind' vertrouwen hebben in de uitkomsten van RCT's? De genoemde funderende principes mogen dan eenvoudig lijken, bij de praktische uitwerking kunnen de problemen groot zijn en daarover is zeer veel gepubliceerd.[5] Methodologen kunnen verschillen in hun opvattingen inzake de aanpak van diverse aspecten van een RCT. In de tijd gezien zijn er bij een RCT steeds een aantal achtereenvolgende stadia: het ontwikkelen van een idee om te gaan toetsen, het opzetten en uitvoeren van de RCT, de procesmatige en inhoudelijke bewaking tijdens de uitvoering (monitoring), de bewerking en rapportage van de resultaten en ten slotte de interpretatie daarvan. De normatieve beoordeling van deze probiotica-RCT spitst zich toe op twee van deze onderdelen: (1) was het een goed idee om het effect van probiotica bij patiënten met ernstige pancreatitis te gaan onderzoeken, en (2) was er een deugdelijke voortgangsbewaking van dit onderzoek? Aangezien er 15 patiënten meer stierven in de behandelde groep, lijkt een 'nee' als antwoord op beide vragen voor de hand te liggen. Te simpel? Ja! Daarmee komen we op de vraag naar de methode van onderzoek om in letterlijke en figuurlijke zin recht te doen aan de trialsituatie.

3. De kundigheid van goed calamiteitenonderzoek

Niets is zo gemakkelijk als wijs te zijn ná de gebeurtenis, niets is zo moeilijk als de juiste beslissing te nemen in situaties van onzekerheid. Wie, geïnformeerd over de afloop, terugkijkt naar de gebeurtenissen voorafgaande aan een rampspoed ziet het onheil al aankomen. De Israëlische psycholoog Baruch Fischhoff verrichtte begin jaren zeventig van de vorige eeuw onderzoek naar de effecten van wijsheid achteraf aan de hand van een aantal experimenten. Die maakten duidelijk dat retrospectieve beoordeling mét kennis van de afloop bij de beoordelaar resulteerde in een toename van de mate waarin de gebeurtenis als voorspelbaar werd ervaren en in vereenvoudiging van causale relaties tussen het handelen en de schadelijke uitkomst.[6] Met kennis van de afloop oordeelt men ook al snel dat degenen die het onheil veroorzaakten niet redelijk en zorgvuldig hebben gehandeld.[7]

De implicaties van kennis van de onfortuinlijke uitkomst zijn enerzijds dat daarmee meer schuld in de schoenen van de verantwoordelijken wordt geschoven en anderzijds dat het moeilijk is om van zijn fouten te leren omdat gekleurde conclusies worden getrokken. Tal van observaties en vooral simulaties hebben duidelijk gemaakt dat er bij juridische beoordeling grote verschillen bestaan tussen beslissers met en beslissers zonder voorkennis van de afloop.[8]

Juist wanneer er normatief – en dus ook objectief – geoordeeld moet worden over de gedragingen van een individu of een groep rijst de vraag of het mogelijk is die 'hindsight bias' buiten spel te zetten. Concreter: hoe dien je bij deze probiotica-trial de gang van zaken te onderzoeken zonder je te laten beïnvloeden door het gegeven van de oversterfte? Vanuit de wens om uit gemaakte 'fouten' werkelijk lering te kunnen trekken, is een methodologie ontwikkeld die niet alleen toepasbaar is voor calamiteitenonderzoek in technische maar ook in juridische zin.[9] De belangrijkste principes daarbij zijn:

- men reconstrueert stap voor stap het proces vanaf de beginfase met de oorspronkelijke richting van de tijd mee (prospectief in plaats van retrospectief);
- men reconstrueert het proces als deelnemer binnen de structuren en systemen waarbinnen de gebeurtenissen plaatsvonden (als 'binnenstaander' in plaats van als buitenstaander);
- men onderzoekt welke *ex ante* geformuleerde voorschriften of regels er in die situatie van toepassing zouden kunnen zijn;
- men ijkt het handelen aan de vigerende *ex ante* geformuleerde voorschriften of regels.

De negatieve afloop blijft bij deze wijze van onderzoek dus steeds buiten beeld, omdat die op keuzemomenten eenvoudig niet bekend kon zijn. Of je je nu als methodoloog, statisticus of als juridische beslisser over deze zaak gaat buigen, steeds zijn deze regels bepalend voor de wijze waarop een kwestie als die van de probiotica-trial onderzocht dient te worden. 'Debiasing' helpt voorkomen dat (onbewust) naar een resultaat wordt toegewerkt waarmee de calamiteit kan worden verklaard. Als ik het artikel van Frijters en Van Houte of het stuk van Gill lees, zijn die allemaal in de val van de 'hindsight bias' gelopen en hebben geen debiasing-technieken willen toepassen. Ik keer terug naar de in paragraaf 2 al eerder genoemde twee kernvragen en wil die proberen te beantwoorden op de wenselijke wijze als geschetst. Daarbij zijn ten minste vier onderling samenhangende dimensies te onderscheiden: een medische, een ethische, een methodologische en een juridische.

4. Relevantie van de vraagstelling en juistheid van gekozen methode van onderzoek

Wanneer artsen een idee voor een onderzoek krijgen en die vraagstelling willen gaan uitwerken, zal minstens aan de volgende voorwaarden voldaan dienen te worden: er is een zinvolle medische rechtvaardiging voor de studie, die proefneming is ethisch verantwoord, praktisch uitvoerbaar en op methodologisch juiste wijze ontworpen en uitgevoerd. Hierop is regelgeving van toepassing in de Wet medisch-wetenschappelijk onderzoek met mensen (hierna: WMO). In artikel 3 van deze WMO worden in lid a-j inhoudelijke en organisatorische

5. Zie voor een goed overzicht S.P. Glasser & G. Howard, 'Clinical trial design issues: at least 10 things you should look for in clinical trials', *Journal of Clinical Pharmacology* (46) 2006, p. 1106-1115.
6. B. Fischhoff, 'Hindsight ≠ foresight: the effect of outcome knowledge on judgment under uncertainty', *Journal of Experimental Psychology: Human Perception and Performance* (1) 1975, p. 288-299.
7. M. Kelman et al., 'Decomposing hindsight bias', *Journal of Risk and Uncertainty* (16) 1998, p. 251-269.
8. E.M. Harley, 'Hindsight bias in legal decision making', *Social Cognition* (25) 2007-1, p. 48-63. De literatuurlijst biedt tal van voorbeelden.
9. Een goed overzicht daartoe biedt het boek van Sidney Dekker, *The field guide to understanding human error,* Aldershot: Ashgate 2006.

eisen aan het wetenschappelijk onderzoek gesteld. Die worden getoetst door een onafhankelijke commissie.
Bij het opzetten van deze RCT om te onderzoeken of probiotica de kans op infectie bij ernstige pancreatitis konden verminderen, was het frequent optreden van slecht te behandelen infecties de medische rechtvaardiging ervan. De logica van de vraagstelling werd tevoren onderbouwd.[10] Dat de uitkomst van de studie anders bleek dan verondersteld, vormt geen overtuigend argument om de onderzoekers te verwijten dat ze een verkeerde keuze hebben gemaakt met alle nadelige gevolgen van dien. Dat is écht wijsheid achteraf. Het komt geregeld voor dat de verwachtingen waarmee een RCT begonnen wordt precies in het tegendeel blijken uit te monden.
Hierboven noemde ik al als voorbeeld hormoonbehandeling in de overgang, waarbij als heilzaam beschouwde hormonale interventie pas bij een RCT schadelijk bleek. Een ander voorbeeld is het geven van hoge doses corticosteroïden (ontstekingsremmende bijnierschorshormonen) bij hersenletsel door een ongeluk in de CRASH-trial.[11] Ook hier was het effect anders dan verwacht: er stierven meer mensen in de behandelde groep. Dat is nu juist de kracht van een RCT: het is immers de beste methode om de waarheid boven tafel te krijgen.
Een ander punt van commentaar is de wijze waarop toestemming werd verkregen van de patiënten voor hun deelname[12] aan het onderzoek en de plicht om deelnemers te waarschuwen als zich tijdens het onderzoek nadelige effecten openbaren waarbij participanten de mogelijkheid dienen te hebben hun deelname aan het onderzoek te beëindigen.[13] Bij *informed consent*, of die nu voor reguliere of voor experimentele behandeling moet worden verkregen, blijken de juridische ideeën en de reële praktijk ver uit elkaar te liggen. Deze problematiek is complex.[14] Het opvoeren van dit punt lijkt ook in belangrijke mate te zijn ingegeven door de ongunstige afloop van de studie. De beoordeling of hier op de juiste wijze voorstelling van zaken werd gegeven aan patiënten of hun vertegenwoordigers dient uit te gaan van het *ex ante* perspectief. Wel wijzen Frijters en Van Houte terecht op het probleem hoe geïnformeerde toestemming te verkrijgen bij ernstig zieke patiënten.
Wellicht ten overvloede: wie weet heeft van de uitkomsten van de probiotica-studie ziet het onheil aankomen en vraagt zich bijgevolg af of de proefpersonen hiertegen niet beter hadden moeten worden beschermd. De vragen of de trial zinvol was en adequaat door de medisch-ethische commissie beoordeeld werd en of de rekrutering van deelnemers zorgvuldig genoeg is geschied, zijn natuurlijk gerechtvaardigd. De wijze waarop naar het antwoord zal worden gezocht, vraagt echter een aanpak zoals in paragraaf 3 beschreven.

5. Bewaking van de veiligheid van de deelnemers

De critici van het probiotica-onderzoek richten hun bezwaren vooral tegen het werk van de 'data monitoring committee' (hierna: DMC). In hun ogen heeft die DMC de kans voorbij laten gaan tijdig te ontdekken dat er in de behandelde groep meer personen overleden, omdat daarvoor de verkeerde werkwijze (drievoudige blindering) en de verkeerde statistische benadering werden gehanteerd. Met name bij dit onderwerp is het essentieel de valkuil van wijsheid achteraf te vermijden. Wat zijn de functies van de DMC en langs welke weg worden die gerealiseerd? Een goed overzicht daarover, geschreven naar aanleiding van deze zaak, is dat van klinisch-epidemioloog Tijssen.[15]
Kenmerkend voor dit soort RCT-studies is dat tevoren bepaald wordt hoe groot de onderzoeksgroep dient te worden om overtuigende conclusies te kunnen trekken. Daarna gaat men van start en worden zich aandienende patiënten opgenomen in de studie. Het onderzoek heeft daarmee een sequentieel karakter: in de loop van maanden of jaren stromen de deelnemende proefpersonen binnen en hun gegevens worden geregistreerd, verwerkt en geanalyseerd. De bewaking door de DMC verloopt periodiek: met tussenpozen worden de resultaten van het onderzoek beoordeeld.
De eerste en belangrijkste functie van de DMC is die van bewaking van de veiligheid, de tweede de tussentijdse analyse van de werkzaamheid. Steeds zijn de algemene vragen: kan het onderzoek ongewijzigd worden voortgezet, dienen er aanpassingen te worden gedaan of moet het onderzoek worden gestaakt? Een trial kan door de DMC tussentijds gestopt worden omdat (1) er in de behandelde groep te veel ernstige bijwerkingen – bovenal sterfte – optreden, (2) er al duidelijk is geworden dat het middel zo effectief is dat voortzetting onnodig is en de therapie per direct algemeen beschikbaar gesteld kan worden[16] en (3) als voortzetting zinloos is omdat de interventie geen enkel positief effect toont. Dit zijn drie essentieel verschillende situaties die daarom ook steeds een verschillende benadering vergen. Wat is de rol van statistische criteria bij de besluitvorming over voortijdige beëindiging? Daarover is al veel gepubliceerd, maar duidelijk is dat er bij statistici verschillende opvattingen bestaan met welke methoden moet worden berekend wanneer de *stopping rule* in werking zou moeten treden.[17] Het is gepast om in dit verband de VU-hoogleraar kansrekening Ronald Meester te citeren[18]:

10. H.C. van Santvoort et al., 'Potentiële rol voor probiotica bij de preventie van infectieuze complicaties tijdens acute pancreatitis', *Nederlands Tijdschrift voor Geneeskunde* (150) 2006, p. 535-540.
11. *The Lancet* (364) 2004, p. 1321-1328.
12. Art. 6 lid 5-9 WMO.
13. Art. 10 lid 1 WMO.
14. R.W.M. Giard, 'Informed consent: van juridische theorie naar medische praktijk – en weer terug!', in: W.H. van Boom, I. Giesen & A.J. Verheij (red.), *Gedrag en privaatrecht. Over gedragspresumpties en gedragseffecten bij privaatrechtelijke leerstukken*, Den Haag: Boom Juridische uitgevers 2008, p. 103-129. Zie ook M.C. de Vries & E. van Leeuwen, 'Ethiek van medisch-wetenschappelijk onderzoek: informed consent en de therapeutische misconceptie', *Nederlands Tijdschrift voor Geneeskunde* (152) 2008, p. 679-683.
15. J.G.P. Tijssen, 'Bewaking van de veiligheid van deelnemers aan gerandomiseerd klinisch onderzoek: grondslagen en methoden', *Nederlands Tijdschrift voor Geneeskunde* (152) 2008-12, p. 674-678.
16. Dat is bijvoorbeeld gebeurd bij geneesmiddelen die de virusproductie bij hiv-infecties onderdrukken.
17. Zie bijvoorbeeld S. Todd et al., 'Interim analyses and sequential designs in phase III studies', *British Journal of Clinical Pharmacology* (51) 2001, p. 394-399 en S.J. Pocock, 'Current controversies in data monitoring for clinical trials', *Clinical Trials* (3) 2006 p. 513-521.
18. R. Meester, 'Statistiek en kansrekening in het strafrecht', *Ars Aequi* 2007-9, p. 675-677.

'Statistische berekeningen leveren getallen op en die getallen lijken daardoor objectief. Dat zijn ze echter geenszins, vanwege het feit dat de keuze van het te gebruiken model meestal helemaal niet vastligt, en onontkoombaar het gevolg is van persoonlijke voorkeuren van de statisticus. Bij de keuze van het model worden veel individuele beslissingen genomen, en dat feit alleen maakt dat elke uitkomst relatief is. Deze constatering moet niet als kritiek gezien worden, maar als een onontkoombare realiteit waarmee rekening dient te worden gehouden.'[19]

Als de DMC zich bij de tweede interim-analyse gaat buigen over de sterftecijfers in de twee groepen van het onderzoek blijkt er verschil, zowel in absolute aantallen als in proporties. Zie de tabel:

Groep	Overleden (%)	Levend (%)	Totaal (%)
Probioticagroep	14 (14,9%)	80 (85,1%)	94 (100%)
Placebogroep	6 (6,7%)	84 (93,3%)	90 (100%)

De sterfte in de behandelde groep is zowel in absolute aantallen als in percentage hoger dan in de placebogroep. Hoe kunnen we deze verschillen verklaren? Is dit op basis van toeval of een schadelijk effect van de behandeling? Welnu, er zijn ten minste al drie verschillende soorten rekenmodellen denkbaar voor de statistische analyse van deze gegevens in de tabel.
In de eerste plaats de weg van *hypothesetoetsing*, de 'klassieke' statistische significantietoetsing. Er worden twee veronderstellingen tegenover elkaar geplaatst: de nulhypothese (er is géén werkelijk verschil in sterfte tussen de groepen) en de alternatieve hypothese (er is wél sterfteverschil). De hypothese die primair getoetst wordt, is de nulhypothese. Bepaald wordt hoe waarschijnlijk het gevonden verschil is als de nulhypothese waar is. Is die kans kleiner dan een vooraf bepaalde grens, veelal 5%, dan wordt de nulhypothese verworpen. Is die kans groter – en dat was hier het geval – dan wordt de nulhypothese niet verworpen. Wil dat dan zeggen dat oversterfte onmogelijk is? Neen, die mogelijkheid blijft aanwezig. Een belangrijke vraag is: hoe groot is de kans op dit verschil gegeven dat de nulhypothese als juist wordt beschouwd?
De berekening van statistische significantie kan op twee manieren geschieden, namelijk door eenzijdige of tweezijdige berekening van de p-waarde. Voor de niet-ingewijden in de statistiek is dit abracadabra, maar voor dit betoog is relevant dat we hier te maken hebben met het kiezen van een model. Die keuze tussen een- of tweezijdig toetsen wordt bepaald door het feit of door toeval bepaalde afwijkingen van het sterftecijfer zowel naar boven als naar beneden kunnen uitschieten of dat dit eigenlijk maar één kant op kan. In het laatste geval valt de keuze op eenzijdig toetsen. Gill, in diens kielzog Frijters en Van Houte en ook de Maastrichtse statisticus Schouten[20] stellen dat hier de eenzijdige toetsing op zijn plaats was. Die benadering zou duidelijk gemaakt hebben dat er een betekenisvol verschil in sterfte was en dus een argument hebben verschaft om de trial te stoppen. Bij tweezijdig toetsen, de in de trial gehanteerde methode, bestond géén statistisch significant verschil. Maar is het logisch te veronderstellen dat de sterftekans maar naar één kant kon uitschieten? Is dat niet vooral achteraf ingegeven door de oversterfte? In de statistische leerboeken wordt aangegeven dat eenzijdig toetsen slechts bij wijze van uitzondering wordt toegepast. De vraag wanneer wél eenzijdig getoetst kan worden, levert onder statistici géén eenduidig antwoord op. Ook bestaat er discussie welke p-waardegrens tevoren dient te worden gekozen als drempel. Dergelijke overwegingen dienen vooraf aan het onderzoek te worden gemaakt.
Een tweede benadering, anders dan die van het testen van hypothesen, is die van *schatting van betrouwbaarheidsintervallen* (confidence testing[21]). Het verschil in sterfte bedraagt in het probiotica-onderzoek 8%, maar dat is een schatting gebaseerd op deze populatie. Als je dit experiment zeer vele malen zou herhalen, zou er dan steeds dezelfde uitkomst zijn? Met een rekenmodel is te becijferen tussen welke waarden het sterfteverschil bij 95% van de experimenten zou liggen. Voor onze groep is dat 95%-betrouwbaarheidsinterval minus 0,9 en plus 17,5%. Omdat dit interval ligt tussen een waarde onder en een waarde boven nul, bestaat geen steun voor de veronderstelling dat hier sprake is van oversterfte (als het interval bijvoorbeeld tussen 4% en 17% had gelegen was dat wel het geval). Met deze methode zou de DMC ook tot voortzetting hebben besloten.
Een derde aanpak is het gebruik van *Bayesiaanse waarschijnlijkheidsrekening*, een methode die de laatste jaren in opkomst is.[22] Tijdens het verloop van de trial kunnen kansen op effecten en neveneffecten (bijv. sterfte) worden geschat. Ook hier spelen aannames over de samenstelling van de populatie en kansen op gebeurtenissen een belangrijke rol. Bij gebrek aan voldoende beschikbare gegevens in de publicatie van Besselink e.a. kan deze benadering (nog) niet worden toegepast. Het is echter zeer de moeite waard deze benadering achteraf op deze probiotica-groep te toetsen om te zien tot welk resultaat dit zou hebben geleid.
Bij een tussentijdse balans hebben we te maken met vooralsnog kleine aantallen patiënten. Naarmate de steekproefgrootte stijgt, neemt de statistische onzekerheid af. De vraag bij interimanalyse is epistemologisch van aard: wat zouden we willen weten en wat kúnnen we weten? De oplossing van dit probleem ligt niet in de juiste keuze van de statistische methode. Een effect dat zich nog niet duidelijk openbaarde tijdens de studie kan aan het einde ervan, bij inmiddels voldoende omvang van de studiegroep, wel blijken. Wat ook duidelijk wordt, is dat verschillende mogelijkheden voor statistische afleidingen tot uiteenlopende conclusies kunnen leiden. Naast de kwestie welke statistische methode we in de gegeven situatie zouden moeten kiezen is er nog een

19. P. 676.
20. 'Statistici: sterfte probiotica onnodig hoog', *NRC Handelsblad* 8 mei 2008.
21. D.G. Altman et al., *Statistics with confidence*, BMJ Books 2000.
22. A. Kpozèhouen et al., 'Use of a Bayesian approach to decide when to stop a therapeutic trial', *American Journal of Epidemiology* (161) 2005, p. 595-603 en D.A. Berry, 'Bayesian clinicl trials', *Nature Reviews Drug Discovery* (5) 2006, p. 27-36.

andere: hoe interpreteer je de uitkomsten?[23] Getallen mogen objectief lijken – zie nogmaals het citaat van Meester – we kunnen erin zien wat we willen zien. Dat is het gevaar van *confirmation bias*.
Een ander punt van kritiek van Frijters en Van Houte is dat bij deze studie sprake was van een drievoudige blindering: de leden van de DMC wisten niet welke groep wel en welke niet de behandeling kreeg. Er zijn goede argumenten om voor deze drievoudige blindering te kiezen, maar er bestaat tussen methodologen desalniettemin – alweer! – verschil van mening over die wenselijkheid.[24] Achteraf bezien is het makkelijk kiezen.
In alle publicaties over het wel of niet stoppen van een trial wordt gesteld dat die besluitvorming complex is en dat statistische argumenten een belangrijke rol spelen maar zeker niet allesbepalend zijn. Niet tijdig stoppen kan leiden tot onnodige schade bij de proefpersonen, maar ook bij het op onjuiste gronden beëindigen van een trial kan schade worden aangericht, zelfs overlijdensschade, omdat toekomstige patiënten ten onrechte een behandeling wordt onthouden. De afwegingen die een DMC moet maken zijn dus complex en daardoor kunnen ze het eerder fout dan goed doen. Daarnaast bestaan er verschillende soorten inzichten en welke is dan vanuit ex ante perspectief de beste?

6. Conclusies

Wanneer we terugkijken op een wetenschappelijk onderzoek waarbij onverwacht veel proefpersonen overleden, wordt er dus veel van ons gevraagd bij het pogen daarover objectief te oordelen. Dát een dergelijke uitkomst ten minste om een kritische beschouwing vraagt, is vanzelfsprekend; een dergelijke oversterfte is geen trivialiteit. Of we nu vanuit wetenschappelijk of vanuit maatschappelijk perspectief een oordeel willen vormen, we zullen ons bewust dienen te zijn van mogelijke dwaalwegen. Maar al te gemakkelijk belanden we in de valkuil van wijsheid achteraf. Het artikel van Frijters en Van Houte getuigt daar mijns inziens van. Het tegenovergestelde daarvan is de valstrik veel, zo niet alles, te relativeren met een ex ante bril op.
In dit probiotica-debat eisen de statistici een belangrijke rol op. Dat is niet voor het eerst. Bij de strafzaak tegen Lucia de B. was er een heftig dispuut tussen statistici of en zo ja hoe aangetoond kon worden dat er een verband bestond tussen haar aanwezigheid en het overlijden van patiëntjes. Er wordt daarom in toenemende mate aandacht besteed aan de vraag welke rol statistische bewijsvoering zou kunnen spelen bij rechtspraak.[25] Wanneer het tot civiele procedures komt naar aanleiding van deze trial, zal eveneens de vraag naar de betekenis van statistische argumenten rijzen. Wat zijn de mogelijkheden en waar liggen de grenzen van statistiek als normatief hulpmiddel? Dat debat bevindt zich wat mij betreft nog maar in een pril stadium.
In mijn betoog heb ik willen toelichten dat bij het beoordelen van de gang van zaken medische, ethische, methodologische, statistische én natuurlijk juridische elementen integraal een rol spelen. In de besproken kwestie is het in ieders belang dat er zo objectief mogelijk onderzocht wordt of de juiste weg werd bewandeld. Feitelijk hebben de onderzoekers met hun persconferentie daarom gevraagd. Het ideale deskundigenonderzoek zal dan ook zonder wijsheid achteraf alle in dit artikel opgesomde punten moeten verdisconteren.
Uit dit strijdgewoel komt wat mij betreft één belangrijk signaal naar boven. Als er geen sprake zou zijn van verwijtbare onzorgvuldigheid in juridische zin bij de DMC, dan is toch wel gebleken dat de bewakingsmethoden voor de veiligheid niet perfect zijn. De gegevens van deze studie zouden opnieuw met alle bestaande statistische modellen, als hierboven geschetst, kunnen worden geanalyseerd en vergeleken om te onderzoeken of nadelige effecten eerder en beter gesignaleerd kunnen worden. Zowel de preventieve werking van het aansprakelijkheidsrecht als wetenschappelijke reflectie over de gebeurtenissen hebben hetzelfde oogmerk: we willen herhaling van zo'n drama proberen te voorkomen! Als we dit doen en betere wegen vinden, hebben rechtspraak én wetenschapsbeoefening vooruitgang geboekt.

23. T.J. Kaptchuk, 'Effect of interpretive bias on research evidence', *British Medical Journal* (326) 2003, p.1453-1455.
24. J.P. Vandenbroucke, 'Dwalingen in de methodologie XIV. Het voortijdig beëindigen van een gerandomiseerde trial', *Nederlands Tijdschrift voor Geneeskunde* (143) 1999, p. 1305-1308.
25. Zie bijvoorbeeld M.J. Sjerps & J.A. Coster van Voorhout (red.), *Het onzekere bewijs. Gebruik van statistiek en kansrekening in het strafrecht*, Deventer: Kluwer 2005.