De gebeurtenissen rond het onderzoek met probiotica hebben de afgelopen week nogal wat stof doen opwaaien. Bij een klinische studie in het AZU bleek dat van de 300 patiënten die probiotica toegediend hadden gekregen, er 24 waren overleden, terwijl in een controlegroep van ook 300 patiënten het aantal sterfgevallen tot 9 beperkt was gebleven. De uitkomst van dit "dubbelblind" experiment werd in de pers omschreven als een "nachtmerrie". Er bleek echter nog iets aan de hand te zijn. Halverwege het experiment werd volgens de persberichten de tussenbalans door de monitoring groep van deze klinische studie opgemaakt, en toen bleek dat het verschil tussen testgroep en controlegroep zich al aftekende. Getallen hierover zijn niet publiekelijk bekend gemaakt, kennelijk omdat anders een wetenschappelijke publicatie in een "gerenommeerd" tijdschrift in gevaar zou komen. Wat we echter wel te horen kregen, was dat de verschillen halverwege het experiment nog toevallig (in vaktermen "niet significant") zouden kunnen zijn, en dat er dus geen reden was om de rest van het experiment af te blazen. Hiervoor was kennelijk ook een statisticus geraadpleegd, al weten we niet wie dat was. Over deze gang van zaken willen wij enkele opmerkingen plaatsen.

Allereerst moet gesteld worden dat het helemaal niet gezegd is dat de uiteindelijke uitkomst überhaupt een significant verschil aangeeft tussen controle- en testgroep. In technische taal gezegd, ligt het er maar aan wat je als significantieniveau hanteert, en ook nog eens welke test je uitvoert. Als je bijvoorbeeld zou kiezen voor de standaard Fisher test, dan is er geen significant verschil bij een niveau van 1%; dat wil zeggen dat wanneer de verdeling van de sterfgevallen volkomen willekeurig zou zijn, in meer dan 1% van de gevallen er een uitkomst te zien zou zijn die extremer is dan deze. Er is wel een significant verschil bij een niveau van 5%. De keuze voor 5% is standaard maar arbitrair.

Maar nog veel belangrijker is de volgende overweging. We weten uit de diverse persberichten dat halverwege de rit zich al een verschil aftekende tussen controlegroep en testgroep, en dat dit verschil in het *nadeel* was van de testgroep. Het medicijn werkte dus eigenlijk de verkeerde kant op, maar kennelijk nog wel binnen de fluctuaties die binnen het toeval mogelijk zijn. Om het experiment succesvol te laten zijn, was het dus nodig dat in de 2e helft van het experiment deze tendens geheel omgebogen zou worden tot in het *voordeel* van het medicijn, en wel op een zodanig sterke manier dat het voordeel zelfs significant zou zijn. Dit is, in alle redelijkheid, werkelijk teveel gevraagd. Dat is wellicht intuïtief al duidelijk, maar we geven toch een rekenvoorbeeld om onze intuïtie te staven.

Laten we er even vanuit gaan dat de trend halverwege vergelijkbaar was met de uiteindelijke uitkomst. Laten we dus zeggen dat halverwege de rit 4 patiënten uit de controlegroep waren overleden, en 12 patiënten uit de testgroep. Een kleine rekensom leert dat dit verschil inderdaad niet significant is bij een niveau van 5%. Echter, zou deze trend zich alleen maar hebben voortgezet, dan zou een verschil van 8 versus 24 doden *wel* significant zijn, maar nog steeds in de "verkeerde" richting. Om de uitkomst zo om te buigen dat het verschil significant zou worden in het voordeel van de testgroep, zou het verschil in de 2e helft van het experiment werkelijk enorm hebben moeten zijn. Zelfs als we ervan uitgaan dat geen enkele patiënt uit de testgroep zou komen te overlijden gedurende de 2e helft van het experiment, dan zouden er nog 21 extra sterfgevallen in de controlegroep

“nodig” zijn om het verschil significant te maken – bij een niveau van 5% - in het voordeel van de testgroep. De kans hierop zal buitengewoon klein zijn. Ook wanneer de resultaten halverwege bijvoorbeeld 6 tegen 9 waren is deze redenering geldig, met net even andere getallen.

We concluderen dat het feit dat halverwege het experiment het verschil tussen de groepen niet significant was, helemaal geen argument had mogen zijn om het experiment voort te zetten. Op dat moment was namelijk al duidelijk dat behoudens zeer extreme gebeurtenissen, het toedienen van de probiotica geen significant voordeel meer zou kunnen opleveren aan het eind van het experiment. Onbegrip van de wereld van kansen en statistiek in een monitoring groep heeft hier dus wel zeer ernstige gevolgen gehad. De statisticus die halverwege geraadpleegd werd valt mogelijk helemaal niets te verwijten: als hem/haar alleen gevraagd is of de getallen een significant verschil aangaven zonder begeleidende context, dan heeft hij/zij correct antwoord gegeven. Vaak is bij statistische overwegingen het stellen van de juiste vraag en het duidelijk maken van de context minstens zo belangrijk als de statistische berekeningen zelf; statistiek is soms meer filosofie dan wiskunde.

Wrang is ook dat de ambitie om zelfs onder deze omstandigheden nog steeds te publiceren hier opening van zaken tegenwerkt. Het publiek heeft er recht op te weten wie verantwoordelijk was voor de gemaakte keuzes, en welk cijfermateriaal daaraan ten grondslag lag.

Prof.dr. Ronald Meester (hoogleraar waarschijnlijkheidsrekening aan de Vrije Universiteit)
Dr. Pieter ter Steeg (Ter SteegMC, Microbiologische en Maatschappelijke Risicoweging, Gouda, voormalig wetenschapsleider Voedingsmicrobiologie Unilever)