

LIES, DAMNED LIES AND STATISTICS

De uitdrukking *lies, damned lies and statistics* zou volgens de Amerikaanse schrijver Mark Twain voor het eerst door de negentiende eeuwse Engelse premier Benjamin Disraeli (1804–1881) zijn gebezigd, en wel als volgt: “There are three kinds of lies: lies, damned lies and statistics”. Deze quote is voor het eerst op papier verschenen in de Engelse krant ‘The National Observer’ van 8 juni 1891: ‘Sir, – It has been wittily remarked that there are three kinds of falsehood: the first is a ‘fib’, the second is a downright lie, and the third and most aggravated is statistics’. Deze uitdrukking was overigens, zoals bleek uit latere uitgaven van dezelfde krant, toen al lang gemeengoed in het spraakgebruik (overigens met dank aan onder andere de Engelstalige Wikipedia). Echt goed toegepaste statistiek is evenwel helemaal geen leugen en in zijn uitkomst onweerlegbaar. Het probleem is echter dat statistiek in ‘wetenschappelijke’ medische publicaties vaak slecht en soms bewust verkeerd wordt toegepast.

Hoe vaak deugt de statistiek niet in medische publicaties?

Aan de vooravond van het congres ‘Advanced Clinical Nutrition 2011’ zat ik in Scheveningen te dineren met enkele collega’s, waaronder de natuurkundige/arts Paul R. Cheney, die gepromoveerd is op een onderwerp betreffende de systematische fout en tegenwoordig onderzoek doet naar het chronisch vermoeidheidssyndroom door meting van de Gibbs-energie van het hart. We kwamen toen onder meer op de statistiek in medische tijdschriften die, naar we beiden vonden, vaak niet in orde is. Cheney vroeg mij in hoeveel procent van de medische publicaties de statistiek volgens mij niet klopt. Afgaande op mijn geheugen antwoordde ik ongeveer 90%. Zijn (chagerende) antwoord luidde “one hundred percent”. Toen ik wat verbaasd keek bij deze reactie herhaalde hij zijn antwoord. Later bedacht ik dat als het achterwege laten van correcties voor systematische fouten wordt meegerekend je natuurlijk hoger uitkomt dan schattingen die alleen op een analyse van de toegepaste statistische test zijn gebaseerd, maar 100% moet gechargeerd zijn. Dit geeft echter wel aan hoe slecht fysici de statistiek in medische artikelen vinden!

Een klein onderzoekje van mijn kant leverde later het volgende op: in het ‘British Journal of Psychiatry’ bleek in 45% van de artikelen na publicatie de statistiek niet in orde (White SJ; B. J. Psych. 135:336-342, 1979). Een ander onderzoek uit ‘eigen’ stal, nu bij het ‘British Medical Journal’ leverde achteraf (!) fouten op in 32 van 61 artikelen. Bij 5 artikelen moesten de hoofdconclusies zelfs worden ingetrokken (wel BMJ, uw reviewers waren dus gewoon slecht), zie Gore SM et al; BMJ 1:85–87, 1977. De frequentie van statistische fouten in medische publicaties wordt op het internet geschat op 30–90% (kennelijk had ik de meest negatieve schatting onthouden). Gemiddeld 60% foutieve statistiek lijkt een redelijke schatting omdat het percentage in de betere medische tijdschriften al ongeveer 50 is. In deze schattingen zijn systematische fouten echter hooguit deels meegenomen. De ‘schatting’ van Cheney was dus inderdaad flink gechargeerd. Hoe dan ook: met deze inleiding heeft de titel van deze column in ieder geval al flink aan legitimiteit gewonnen.

Wat zijn systematische fouten?

Een systematische fout is bijvoorbeeld: je test een geneesmiddel versus een placebo, maar in de controlegroep bevindt 50% zich in een ernstiger ziektestadium, terwijl dat in de experimentele groep 40% is. De laatste groep laat dan sowieso een beter resultaat zien. Ook bij juiste toepassing van de statistiek op de einddata is het onderzoek dan vertekend. Men had moeten stratificeren, dat wil zeggen per stadium moeten loten. Dan was de verdeling over de twee groepen gelijk geweest. Zo’n fout kan ook ontstaan indien de ene waarnemer de experimentele groep doormet en een andere waarnemer de controlegroep, terwijl ze beter elk de helft van beide groepen hadden kunnen doormeten. Etc. etc.

Enkele voorbeelden van verkeerde methodologie/statistiek

Een bekende truc is het doen van eenzijdige toetsing: je vraagt niet of er tussen A en B een verschil is, maar je vraagt slechts of A beter is dan B (je sluit dan de mogelijkheid dat B beter is dan A bij voorbaat uit). Dit leidt zonder dat er van een echt andere situatie sprake is tot een lagere p-waarde (zie Wijvekate ML: ‘Verklarende statistiek’; 13e druk, 1972). Nog een voorbeeld: een toets die een normale (dus symmetrische) verdeling vereist, mag alleen worden toegepast als de te testen data in principe normaal zijn verdeeld. Wil je een dergelijke toets toepassen dan moet je met een toets voor normaliteit eerst vaststellen of de data per groep inderdaad normaal zijn verdeeld, anders kun je beter een verdelingsonafhankelijke toets doen (zie weer Wijvekate). Bij kleine aantallen is dit nog meer van belang! Het klassieke voorbeeld van door systematische fouten gedevalueerd onderzoek is de ATBC-studie: de rokers uit de bèta-caroteengroep waren gemiddeld een paar maanden ouder en ettelijke maanden eerder gaan roken. Omdat de kans op longkanker bij dergelijke mensen sneller dan lineair stijgt, komt in de groep die door de twee eerdergenoemde fenomenen al verder op de longkankercurve zit automatisch na verloop van tijd meer longkanker voor (een van de grootste fouten was hier dat niet eerst op leeftijd was gestratificeerd). Als je vooraf stratificeren en/of achteraf multivariate analyse zou eisen (natuurlijk zou dat moeten!) om te zorgen dat de groepen zo goed mogelijk vergelijkbaar zijn, dan zou het percentage onderzoeken met onjuiste statistiek fors omlaag gaan.

Een ander goed voorbeeld van een ernstige systematische fout is het geruchtmakende onderzoek naar de rol van probiotica bij pancreatitis bij het UMCU (Universitair Medisch Centrum Utrecht), zie het TvOG van juni en augustus 2008. Dat was voor mij scoren voor een open doel!

Zelf heb ik drie keer kort na publicatie van een eigen artikel mijn statistiek verbeterd (een keer gaf dat een verandering in mijn voordeel en twee keer een verandering in mijn nadeel; dit leidde in evenwel in geen van de drie gevallen tot wijziging van de conclusies). Tien jaar geleden heb ik eerst 'Fysisch experimenteren' van G.L. Squires doorgewerkt (inder-tijd het leerboek statistiek voor studenten natuurkunde aan de RU te Utrecht), dat meer dan dertig jaar geleden al werd gebruikt (en nog steeds hoorde ik). Dit vergrootte mijn inzicht zeer, voor iedereen aan te bevelen.

Het visolieonderzoek: statistiek deugt niet!

Een zeer recent voorbeeld van foutieve statistiek is het onderzoek van cisplatinum en dergelijke in relatie tot visolie, eveneens van het UMCU (zie Roodhart JML et al; Cancer Cell 20:370–383, 2011, met in de slipstream van de auteurs ook oncoloog Voest). Daar werd in een onderzoek met slechts 6 muizen per groep (een blancogroep, een cisplatinum-groep en twee verschillende groepen met elk een soort visolie plus cisplatinum) gekeken of visolie invloed kon hebben op de werking van cisplatinum. Met een t-toets werd een minder goede werking van de cisplatinum in de twee afzonderlijke visoliegroepen vastgesteld. Het was evenwel niet duidelijk of er wel **vooraf** een toets was gedaan die aantoonde dat van normale verdelingen sprake was. Ook was niet aangegeven of er een- dan wel tweezijdig is getoetst (het moet natuurlijk tweezijdig). Er is evenwel nog een veel ernstiger probleem: bij kleine aantallen, zoals hier met slechts 6 muizen per groep, is er sprake van een grote fout in de standaarddeviatie. Bij $n = 6$ bedraagt deze maar liefst 31,5% (zie het eerdergenoemde boek van GL Squires, blz. 36–37 en 228–229). In formule voor niet te kleine n : $1/\sqrt{(2n-2)} \times 100\%$. Dit ondermijnt de significantie in hoge mate. Niet voor niets gaf mijn toenmalige docent 'fysische meetmethoden', de hoogleraar fysica Hoogenboom, in 1972–1973 het advies bij aantallen kleiner dan 10 liever geen statistiek te bedrijven. De al genoemde Squires, eertijds professor in Cambridge, stelt algemeen aan de hand van $n = 9$ (de fout in de standaarddeviatie is dan wat kleiner dan bij $n = 6$): 'Het resultaat is een duidelijke waarschuwing tegen te nauwkeurige foutenberekening. We zien bijvoorbeeld dat voor $n = 9$ –toch een niet onbelangrijk aantal metingen– de fout zelf nog een onzekerheid van 25% heeft'. Daarbij komt dat je met twee soorten visolie ook nog eens sneller door toeval significantie kan bereiken. Bij mijn weten betrof het ook nog eens slechts één experiment en werd de cisplatinum bovendien intraperitoneaal en niet intraveneus toegediend. Daarenboven heb ik nog systematische fouten gesignaleerd die de conclusies van het onderzoek verder ondergraven. Ik heb mijn bevindingen bij de onderzoekers te berde gebracht en aangegeven dat een verdelings-vrije toets (bijvoorbeeld Wilcoxon) slimmer was geweest, al denk ik dat ook bij Wilcoxon bij kleine aantallen rekening moet worden gehouden met onzekerheid in de ligging van de verdeling. Verder heb ik de onderzoekers nog gewezen op een recent gerandomiseerd onderzoek (PMID 21328326) bij 31 versus 15 patiënten die wegens kanker respectievelijk cisplatinum en cisplatinum met dagelijks visolie (met 2,5 g EPA+DHA) kregen. Met visolie erbij waren er significant meer regressies (60% versus 25,8%). De precieze samenstelling van de in dit onderzoek gebruikte visolie ken ik niet, maar dit is een van de vele onderzoeken die de conclusies van de UMCU-onderzoekers verder invalideren! Uiteraard kwam dit laatste onderzoek bij de mens in tegenstelling tot het UMCU-onderzoek bij een paar muizen niet in het nieuws! Overigens werkt visolie met vitamine E alleen ook therapeutisch bij kanker (zie PMID 9445198).

Op mijn rapportage aan de Utrechtse onderzoekers, die ik later nog heb gepreciseerd omdat Squires voor mij ook al even geleden is, mocht ik geen reactie ontvangen. Ook ontving ik geen nieuwe statistische analyse met liefst extra raw data. Gezien het ontbreken van een aantoonbaar juiste statistische methode, het direct opzoeken van de publiciteit, maar juist ook vanwege het veelbetekende stilzwijgen bij kritische vragen verdient het Utrechtse onderzoek nu slechts een plaats in de prullenbak!

Opmerkelijk is ook dat de vooral aan KWF Kankerbestrijding te linken hoogleraar 'kanker en voeding', mevrouw Ellen Kampman (zij werd volgens mij niet geheel toevallig zaterdag 8 oktober 2011 door 'de Volkskrant' geïnterviewd) dit visolieonderzoek omarmt. Kampman heeft het onderzoeksrapport dan of niet gelezen of haar kennis van de statistiek is ver ondermaats. Voor het laatste pleit ook het feit dat ze vorig jaar op een symposium een vraag van mij over de eerdergenoemde ATBC-studie handig ontweek. Linksom dan wel rechtsom blundert Kampman door haar steun aan het dubieuze Utrechtse visolieonderzoek te geven, nog afgezien van het feit dat ze het visolieonderzoek bij de mens dat ik al noemde vermoedelijk niet kent (anders had ze dat toch kunnen noemen?).

Conclusie

Lies, damned lies and (onjuiste) statistics zijn derhalve in medisch-wetenschappelijke publicaties, met medeweging van systematische fouten, helaas zeer prevalent, al blijft er gelukkig genoeg gelegenheid om wel goede statistiek te bedrijven, desnoods achteraf tegen de politieke en/of commerciële stroom in. De gang van zaken rondom dit visolieonderzoek heeft wel duidelijk gemaakt dat het de Utrechtse onderzoekers veel gemakkelijker afging om de knop voor de t-toets op de rekenmachine te vinden en de weg naar de schrijvende dagbladvijvers te zoeken dan om tijdens het onderzoek zich in elementaire statistiek te verdiepen c.q. een goede statisticus te raadplegen.