



Universiteit
Leiden
The Netherlands

De ultieme waarheid

Rosendaal, F.R.

Citation

Rosendaal, F. R. (2002). De ultieme waarheid, 304-309. Retrieved from <https://hdl.handle.net/1887/1586>

Version: Not Applicable (or Unknown)

License:

Downloaded from: <https://hdl.handle.net/1887/1586>

Note: To cite this publication please use the final published version (if applicable).

Dwalingen in de methodologie (slot). XXXIX. De ultieme waarheid

F R ROSENDAAL EN L M BOUTER

Steeds wenkend en steeds
wijkend wonderland
(Jan Slauerhoff)

De voorafgaande bijdragen in de artikelenreeks 'Dwalingen in de methodologie' handelden over de juiste aanpak bij een bepaalde onderzoeksvraag, en de bypassen de onderzoeksofzet. Voor een grote verscheidenheid van vraagstellingen op etiologisch, diagnostisch, prognostisch en therapeutisch gebied is besproken hoe deze onderzocht kunnen worden, en wat bij iedere benadering de methodologische valkuilen zijn. Cruciaal hierbij is welke aanpak de kans op een 'waar' antwoord optimaliseert, bij voorkeur met verwijzing naar de kosten wat betreft arbeid, tijd en geld.

Een vraag die het individuele onderzoek ontstijgt, is of het antwoord ook werkelijk waar is. Deze vraag speelt natuurlijk altijd een rol, maar treedt op de voorgrond wanneer er verscheidene studies met verschillende resultaten zijn. Indien de verschillen tussen studies op toevalsfluctuaties berusten, kan in een meta-analyse een precieze gemiddelde conclusie worden berekend. Maar vaak zijn de beschikbare studies heterogeen en helpt een meta-analyse slechts bij het exploreren van de bronnen hiervan door homogene subgroepen te identificeren. Zowel inhoudelijke als methodologische aspecten van de in een meta-analyse betrokken studies kunnen de heterogeniteit van hun conclusies verklaren. Wanneer bijvoorbeeld blijkt dat geblindeerde studies leiden tot andere uitkomsten dan open studies, ligt het voor de hand om vooral aan de geblindeerde studies waarde te hechten. De afgelopen jaren is in meta-analysen aangetoond dat de uitkomsten van een onderzoek (soms) samenhangen met de methodologische kenmerken ervan.^{1,3}

ONDERZOEK MET METHODOLOGISCHE TEKORTKOMINGEN NIET ALTIJD ONWARE RESULTATEN

Colonkanker bij colitis ulcerosa. Het moge duidelijk zijn dat een verkeerd uitgevoerd onderzoek tot onware resultaten kan leiden. Gebrekkig onderzoek leidt echter niet noodzakelijk tot vertekende resultaten. Een voorbeeld is een onderzoek naar het risico op colonkanker bij colitis ulcerosa,⁴ waarbij een van de vragen was of het risico toenam met de duur van de blootstelling. Met an-

SAMENVATTING

– Onderzoek dat niet geheel volgens de regels van de kunst is uitgevoerd, heeft daarom niet altijd tot verkeerde conclusies te leiden. Soms is de vertekening klein, of is het gebrek niet relevant voor de betreffende vraag.

– Evenzeer is het mogelijk dat onderzoek dat perfect is opgezet en uitgevoerd, toch niet het ware antwoord oplevert. Door toeval kunnen bijvoorbeeld onvergelykbare groepen ontstaan.

– Met de regel van Bayes is te begrijpen hoe eenzelfde onderzoeksresultaat bij verschillende deskundigen tot geheel andere conclusies kan leiden.

De ultieme waarheid blijft ongreepbaar.

dere woorden wordt de kans op colonkanker groter als de colitis ulcerosa langer bestaat? Deze voor de praktijk zeer relevante vraag werd onderzocht bij een groep patiënten die werden verwezen met een, soms al vele jaren bestaande, colitis. De patiënten werden na verwijzing prospectief gevolgd. Bij aanvang van de studie werd vastgelegd wanneer de diagnose 'colitis' gesteld was, dit was, omdat het onderzoek in een tertiair verwijzingscentrum werd uitgevoerd, veelal vele jaren eerder gebeurd in een andere praktijk. Hierdoor bevond zich slechts een deel van de follow-up binnen het daadwerkelijke observatievenster van het onderzoek, en een niet onaanzienlijk deel had daarvoor plaatsgevonden. Dit maakte het mogelijk met een observatievenster van 16 jaar uitspraken te doen over een periode van 40 jaar. De observatieduur werd vervolgens ingedeeld in decennia sinds de eerste diagnose van de ziekte, de kans op colonkanker per decennium werd berekend, en men vond inderdaad een sterk stijgende incidentie over deze decennia (figuur 1).

In figuur 1 is de wijze van berekenen van het risico op colonkanker per decennium van ziekteduur weergegeven. Het meerekenen van de ziekteduur voor het tijdsvenster van het onderzoek is theoretisch onjuist en kan leiden tot vertekening. Deze vertekening houdt namelijk een onderschatting van het risico op colonkanker gedurende de eerste decennia in. Dit is eenvoudig in te zien: sommige patiënten die in het onderzoek worden opgenomen, hebben al vele jaren colitis (A en C in figuur 1). Onze blik op de tijd kort na de diagnose van die patiënten is beperkt en er zijn vermoedelijk ook patiënten E en F geweest die tegelijk met de patiënten A en C colitis kregen, doch die nooit in het onderzoek terecht kwamen omdat zich bij hen reeds eerder colonkanker ontwikkelde. In dit onderzoek ontbreekt een duidelijk in de tijd gedefinieerde cohort, en de onderzoekers hebben zich rijk gerekend door follow-up mee te tellen die voor de daadwerkelijke observatieperiode viel. De con-

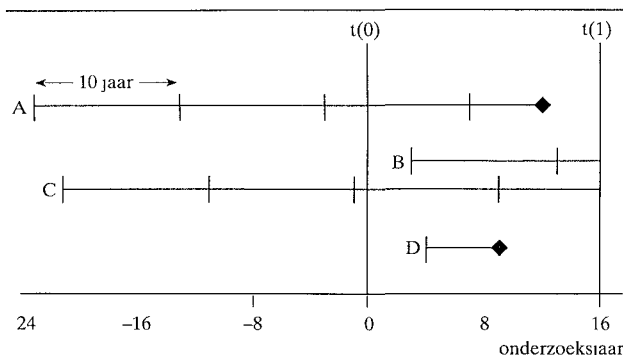
Leids Universitair Medisch Centrum afd Klinische Epidemiologie
Postbus 9600 2300 RC Leiden
Prof dr F R Rosendaal klinisch epidemioloog
VU Medisch Centrum Instituut voor Extramuraal Geneeskundig Onderzoek Amsterdam
Prof dr L M Bouter klinisch epidemioloog
Correspondentieadres: prof dr F R Rosendaal (f.r.rosendaal@lumc.nl)

clusie dat het risico op colonkanker stijgt met de tijd die is verlopen sinds het stellen van de diagnose 'colitis', was dus gebaseerd op vertekend onderzoek. Het betreffende onderzoek is uitvoerig bekritiseerd,⁵ ook in dit tijdschrift,⁶ en is tegenwoordig een dankbaar object in ons methodologieonderwijs. Toen echter enige jaren later een onderzoek werd uitgevoerd met een gelijksoortige vraagstelling, maar met wel een juiste onderzoeksopzet, werden vergelijkbare resultaten gevonden: het risico op colonkanker steeg naarmate de colitis langer bestond.^{7,8}

Veneuze trombose bij bloedgroep AB0 Vaak zal het niet mogelijk zijn om de volledig juiste onderzoeksopzet uit te voeren, bijvoorbeeld omdat de benodigde diagnostische technieken nog niet beschikbaar zijn. Ook in die situaties zijn de tekortkomingen soms niet of nauwelijks van invloed op de waarheid van de resultaten. Een voorbeeld is onderzoek naar de relatie tussen bloedgroep AB0 en het ontstaan van veneuze trombose.⁹ Dit onderzoek werd eind jaren zestig van de afgelopen eeuw voor het eerst uitgevoerd. Toen berustte de diagnostiek van trombose voornamelijk op lichamelijk onderzoek, aangezien er minder mogelijkheden voor objectieve diagnostiek waren dan tegenwoordig (zoals compressie-echografie). Hoewel later onderzoek heeft uitgewezen dat met louter anamnese en lichamelijk onderzoek de diagnostiek van veneuze trombose uiterst onbetrouwbaar kan zijn,^{10,11} had men in deze oudere studies al een aantal voorzorgen genomen. Zo deelde men de patiënten in twee groepen in: degenen bij wie het klinisch beeld zeer overtuigend was en anderen bij wie dat minder het

geval was. De redenering daarbij was dat indien er bias was door overdiagnose, dit vooral in de groep van minder overtuigende gevallen te zien zou zijn. Er bleek echter nauwelijks verschil te bestaan in de risicoschattingen tussen de verschillende groepen. Recent onderzoek bij patiënten met objectief aangetoonde trombose heeft de resultaten uit de oudere studies bevestigd, en inmiddels is duidelijk dat het lagere risico bij bloedgroep 0 komt door de hiermee gepaard gaande lagere bloedspiegels van Von Willebrand-factor en stollingsfactor VIII.¹²

Men moet blijven nadenken Er is dus alle reden om onderzoek met echte of vermeende methodologische tekortkomingen niet te snel naar de prullenmand te verwijzen. Indien men de niet onredelijke stelling aanhangt dat ieder onderzoek tekortkomingen heeft, is dit een geruststellende gedachte. Het methodologisch zwakke onderzoek zal soms tot verwaarloosbaar kleine kwantitatieve verschillen met de ware uitkomst leiden en soms is er zelfs geen enkel effect op de gevonden uitkomst. Dit laatste kan op toeval berusten, of het gevolg zijn van het feit dat het methodologische gebrek er in dat specifieke geval niet toe doet. Soms is dit van tevoren te berekenen: zo leidt het meerekenen van niet daadwerkelijk geobserveerde tijd, zoals bij het voorbeeld van colonkanker bij colitis ulcerosa, alleen tot vertekening indien de betreffende aandoening tot oversterfte leidt (bij colonkanker is dat vanzelfsprekend het geval). De methode is derhalve verdedigbaar indien er geen oversterfte is, en is bijvoorbeeld herhaaldelijk gebruikt bij het vaststellen van het tromboserisico bij individuen met een erfelijke tromboseneiging.^{13,15} De gebruiker van de resultaten van een onderzoek moet zich niet alleen afvragen wat de methodologische tekortkomingen waren en of die in dit geval tot vertekening zullen leiden, maar moet zich ook een beeld proberen te vormen van de vermoedelijke richting van die vertekening. Immers, wanneer er bijvoorbeeld een effect is gevonden en dit vermoedelijk een onderschatting is, dan blijven de kwalitatieve conclusies geldig.



FIGUUR 1 Opzet van prospectief onderzoek naar het risico op colonkanker bij patiënten met colitis ulcerosa.⁴ Hier zijn 4 patiënten (A, B, C, D) weergegeven. Zij worden gevolgd over een tijdsvenster $t(0)$ - $t(1)$, dat 16 jaar omspannt. Sommigen van hen (A, C) zijn bij het begin van de studie reeds bekend wegens colitis ulcerosa; bij anderen (B, D) wordt gedurende de loop van de studie de diagnose gesteld. Bij 2 patiënten (A, D) ontwikkelt zich coloncarcinoom (♦), terwijl de 2 andere patiënten (B, C) de einddatum van de studie bereiken zonder dat zij een coloncarcinoom hebben gekregen. De tijd sinds de diagnose 'colitis ulcerosa' per patient is verdeeld in decennia, en het risico per decennium is het aantal gevallen van coloncarcinoom per decennium gedeeld door het aantal geobserveerde persoonsjaren per decennium. In dit voorbeeld is dit risico $1/35$ voor het 1e decennium ($1/(10 + 10 + 10 + 5)$), 0 voor het 2e en het 3e decennium, en $1/12$ voor het 4e decennium ($1/(7 + 5)$). Het risico is dus in het voorbeeld 2,9 maal zo hoog in het 4e als in het 1e decennium na de diagnose.

METHODOLOGISCH PERFECT ONDERZOEK NIET ALTIJD WARE RESULTATEN

Verder is er onderzoek met onberispelijke methoden. Dit levert helaas toch niet altijd het juiste antwoord op. Daarvoor zijn verscheidene oorzaken aan te wijzen. Ten eerste is het belangrijk de onderzoeksopzet en de gebruikte methoden in de breedste zin te bezien. Het gaat niet alleen om methodologische aspecten, zoals wijze van randomisatie, keuze van de controles of het tijdsvenster van de follow-up. Het gaat ook, en wellicht vooral, om inhoudelijke keuzen die zijn gemaakt in de onderzoeksopzet, zoals eenduidige ziekte-definitie, adequate diagnostiek en zinvolle uitkomstmaten. Het is opvallend dat in het kader van 'evidence-based medicine' lijsten worden opgesteld van zogenaamde 'levels of evidence' die vrijwel volledig op methodologische aspecten gebaseerd zijn, en dat de inhoudelijke keuzen die werden gemaakt daarbij grotendeels worden genegeerd.

Het is van belang om onder ogen te zien dat bepalen van niveaus van bewijskracht om twee redenen feil-

baar zijn. Ten eerste is van nog maar weinig methodologische aspecten empirisch vastgesteld óf en onder welke omstandigheden ze tot vertekening zullen leiden. Ten tweede negeren dergelijke systemen de mogelijkheid dat inhoudelijke keuzen in de onderzoeksopzet eveneens tot vertekening leiden.

Behandeling van rug- en nekpijn. Een mogelijke verklaring voor een onjuist resultaat van een naar deze maatstaven perfect opgezet onderzoek kan zijn dat inadequate maten gebruikt zijn voor vaststelling van blootstelling of ziekte-uitkomst. Een voorbeeld is een gerandomiseerd onderzoek naar de behandeling van rug- en nekpijn.^{16 17} Het meten van verbetering in functioneren en pijnbeleving is niet geheel rechttoe rechtaan, en er werden verscheidene schalen toegepast, waarbij duidelijke en consistente verschillen werden waargenomen bij één van de onderzochte behandelingsstrategieën. Echter, één meetinstrument dat in eerder onderzoek veel was toegepast en dat door sommigen in de literatuur als de standaard en de best gevalideerde schaal werd gezien, bleek bij de patiënten in dit onderzoek al vóór het begin van de therapie nauwelijks enige functiebeperking te laten zien; verbetering op deze schaal was dus niet denkbaar. Was nu deze schaal als enige in het onderzoek toegepast, dan zou er geen enkel effect waargenomen zijn.

Randomisatie is geen garantie. Bij gerandomiseerd onderzoek is het mogelijk dat er ondanks de randomisatie een scheve verdeling van prognostische factoren is ontstaan. Indien dit probleem onbekende determinanten van de bestudeerde uitkomst betreft, is dit vanzelfsprekend niet na te gaan en valt er ook niet voor te corrigeren. Ook bij observationeel onderzoek kunnen belangrijke andere determinanten van de bestudeerde uitkomst ongelijk verdeeld zijn tussen degenen met en zonder de factor die wel bestudeerd wordt (zogenaamde onbekende 'confounding'). Bij onderzoek naar genetische risicofactoren is het mogelijk dat de variant die blijkt samen te hangen met ziekte niet zelf de oorzaak is, maar op zijn beurt weer samenhangt met een andere mutatie (dit heet 'linkage disequilibrium'), of dat de gevonden samenhang berust op een heterogene etnische samenstelling van de onderzoekspopulatie (onbedoelde populatieselectie, ook wel populatiestratificatie genoemd).

Toeval. Tenslotte is er toeval: ook een methodologisch perfect onderzoek kan leiden tot een conclusie in een bepaalde richting doordat geheel toevalligerwijs in de ene groep meer patiënten ziek worden dan in de andere groep. Zo kan het gebeuren dat er een verschil gevonden wordt – tussen therapieën, of tussen risicofactoren – dat er in werkelijkheid niet is, of dat er geen effect gevonden wordt in het onderzoek, terwijl het er in werkelijkheid wel is. Er is geen enkele mogelijkheid in empirisch onderzoek een toevalsbevinding geheel uit te sluiten, zoals reeds in de 18e eeuw door de Engelse filosoof David Hume werd gesteld ('Humes probleem').

Homeopathie. Hoe methodologisch vlekkeloos onderzoek tot onjuiste resultaten kan leiden, laten bijvoorbeeld studies naar het effect van homeopathie fraai zien. Een meta-analyse van gerandomiseerde onderzoeken, zorgvuldig geselecteerd op basis van de hoogste

methodologische standaarden, liet een statistisch significant gunstig effect van homeopathische behandeling zien.^{3 18 19} In dit geval is het naar onze mening onjuist te concluderen dat er overtuigend bewijs is dat deze therapie werkt, ongeacht het hoge niveau van bewijs, aangezien de uitgangshypothese onjuist was: er kan namelijk geen effect zijn van homeopathische geneesmiddelen. Wij geven toe dat dit filosofisch gezien een a-prioristisch standpunt is, maar laten ons leiden door de aan James Oberg toegeschreven waarschuwing: 'Keeping an open mind is a virtue, but not so open that your brains fall out.'²⁰

Vanzelfsprekend is het zo dat indien men bereid is te geloven dat homeopathie een farmacologisch effect zou kunnen hebben, de meta-analyse de juistheid van deze visie ondersteunt. Dit leidt tot de bevreedende conclusie dat de interpretatie van bevindingen uit in methodologische zin perfect onderzoek radicaal kan verschillen. Dit werken wij verderop uit.

CONFLICTERENDE STUDIES

Conflict is regel. Door alle genoemde oorzaken is het niet verrassend dat er over eenzelfde vraagstelling vaak conflicterende studies zijn. Alleen al vanwege de toevalsvariatie is het regel in plaats van uitzondering dat onderzoeken elkaar tegenspreken. Alvorens een onderzoek wordt begonnen, wordt doorgaans aan de hand van het onderscheidingsvermogen (de 'power') berekend hoeveel patiënten geïncludeerd moeten worden. Het onderscheidingsvermogen is de kans dat wanneer een vooraf gespecificeerd effect aanwezig is, dit ook daadwerkelijk in het onderzoek wordt gevonden, en doorgaans wordt hierbij een kans van 80% gehanteerd. Wanneer nu achtereenvolgens twee studies met dit onderscheidingsvermogen van 80% worden uitgevoerd, is de kans dat beide dit bestaande effect zullen aantonen slechts 64%.

Wanneer het niet ongewoon is dat onderzoeken elkaar tegenspreken, dienen we hiermee om te kunnen gaan. In de klassieke statistiek wordt bij een onderzoek een nulhypothese opgesteld ('homeopathie werkt niet') en deze wordt verworpen indien de geobserveerde data de nulhypothese onwaarschijnlijk maken (statistisch significant resultaat). In feite kan wat er gebeurt bij wetenschappelijk onderzoek vergeleken worden met wat er gebeurt bij een diagnostische test:^{21 22} ook hier passen we een imperfecte test toe teneinde een uitspraak over de werkelijkheid te kunnen doen, in dit geval de uitspraak of de patiënt een bepaalde ziekte heeft of niet.

Tabel 1 (bovenste paneel) laat de toepassing van een diagnostische test zien in een populatie waarvan de helft ziek is. Dit gegeven wordt de voorafkans (priorkans of a-priorikans) genoemd (in dit geval 50%). De test heeft goede testkenmerken, zoals een sensitiviteit van 95% (onder 100 zieken is de testuitslag bij 95% afwijkend) en een specificiteit van 99% (onder 100 niet-zieken is de testuitslag bij 99% niet-afwijkend). Wat zegt in dit geval een positieve testuitslag? Dit wordt weergegeven door de voorspellende waarde: van degenen met een positieve testuitslag heeft 99% daadwerkelijk de ziekte. Dit wordt de achterafkans (posteriorikans of a-posteriorikans) genoemd. Het bovenste paneel geeft een scenario

TABEL 1. Voorspellende waarde van een diagnostische test

testuitslag	aantal patiënten		voorspellende waarde van een positieve uitslag
	ziek	niet ziek	
voorbeeld 1			
positief	95	1	95/96 = 99%
negatief	5	99	
totaal	100	100	
voorbeeld 2			
positief	95	100	95/195 = 49%
negatief	5	9 900	
totaal	100	10 000	

dat regelmatig in de spreekkamer zal voorkomen: door anamnese en lichamelijk onderzoek is een voorselectie gedaan alvorens de test wordt uitgevoerd, en de priorkans is daardoor vrij groot geworden.

Anders wordt het wanneer de priorkans kleiner is (onderste paneel): in dit voorbeeld is de prevalentie van de ziekte 1% (100/10.100). In dit geval, bij een test met eenzelfde goede sensitiviteit en specificiteit, heeft minder dan de helft van degenen met een afwijkende testuitslag daadwerkelijk de ziekte (de voorspellende waarde of posteriorkans is 49%). Het merendeel van de patiënten met een positieve testuitslag is in werkelijkheid niet ziek. Het moge duidelijk zijn dat bij screening, dat wil zeggen bij het zoeken naar ziekten met een veelal nog lagere prevalentie, dit probleem van een lage voorspellende waarde van zeer grote betekenis wordt. Overigens heeft ook in het tweede geval de test wel degelijk bijgedragen aan reductie van onzekerheid: de priorkans van 1% stijgt bij een positieve testuitslag tot een achterafkans van 49%, hetgeen een aanzienlijk verschil is.

Overeenkomst tussen een statistische toets en een diagnostische test. We kunnen statistische toetsen beschouwen als diagnostische tests voor de werkelijkheid. Hierbij wordt een iets andere terminologie gebruikt. De gevoeligheid om een effect aan te tonen (sensitiviteit) noemt men, zoals reeds werd vermeld, het onderscheidingsvermogen of de power. Vaak wordt, zoals gezegd, een power van 80% aangehouden. De kans om een effect te vinden dat er in werkelijkheid niet is, wordt het significantieniveau genoemd. Het is vergelijkbaar met: 1 – specificiteit, en wordt bij conventie op 5% gesteld. Op deze wijze bekeken is er ook bij onderzoek een priorkans, en is er bij een 'positief resultaat', in dit geval een significante p-waarde, een voorspellende waarde (posteriorkans) te berekenen. De priorkans voor een diagnostische test is niets anders dan de prevalentie van de aandoening onder degenen met een indicatie voor de betreffende test. Vaak is die prevalentie echter niet bekend en moet de clinicus een intuïtieve schatting maken van de priorkans. Bij statistische toetsen doet zich een analoge situatie voor: de priorkans (op de juistheid van een bepaalde hypothese) kan slechts intuïtief worden ingeschat. Stel dat

de geloofwaardigheid van de hypothese op 50% wordt geschat. Dan zijn we weer bij het bovenste paneel van tabel 1, indien de power 95% was geweest en we een significantieniveau van 1% hadden gehanteerd. Gebruikelijker zijn echter waarden van 80% respectievelijk 5%, en dan krijgen we de getallen zoals in tabel 2.

Bayesiaanse statistiek. Indien men – uitgaande van deze waarden – onderzoek uitvoert met een priorkans van 50%, is bij een significant resultaat de achterafkans 94% (bovenste paneel van tabel 2). Dat percentage is hoog, maar het geeft geen zekerheid, en hoe vaak het onderzoek ook herhaald wordt, nooit zal de kans 100% worden. Wanneer we er op deze manier naar kijken, hetgeen de basis vormt van de zogenaamde Bayesiaanse statistiek, zijn conflicterende studies ineens geen probleem meer. Immers, na een eerste studie, zoals weergegeven in het bovenste paneel van tabel 2, is voor een bepaalde onderzoeksvraag de achterafkans 94% geworden. Deze kan men beschouwen als de vóórafkans voor de volgende studie met deze vraagstelling. Vallen de resultaten daarvan wederom statistisch significant uit, dan zijn we nog zekerder dat er inderdaad een effect is. Valt daarentegen een volgend onderzoek niet statistisch significant uit, dan is de achterafkans wat kleiner geworden (in dit geval 77%, namelijk: $0,20 \times 94 / (0,20 \times 94 + 0,95 \times 6)$), en zijn we iets minder zeker geworden. Terwijl elkaar tegensprekende onderzoeken vaak tegenover elkaar geplaatst worden, waarbij men, doorgaans tevergeefs, probeert vast te stellen welke uitslag 'waar' en welke 'onwaar' is, dragen nu beide bij aan een kwantificering van de 'waarheid'. Men moet zich wel realiseren dat 'waarheid' hier betekent: het vertrouwen in de juistheid van een hypothese na kennisname van de uitslag van de significantietoetsen van de beschikbare studies.

Het theorema van Bayes zoals hier gepresenteerd, maakt ook de evaluatie van positief uitgevallen onderzoek naar homeopathie eenvoudig, zoals weergegeven in het onderste paneel van tabel 2. De voorafkans op een effect van homeopathie is naar onze mening bijzonder klein; daarom is in het voorbeeld een kans van 1/1.000.000 gebruikt. In dat geval geeft ook een positief uitgevallen onderzoek geen enkele aanleiding te menen dat aangetoond is dat homeopathie zou werken, aangezien de ach-

TABEL 2. Voorspellende waarde van een statistische toets

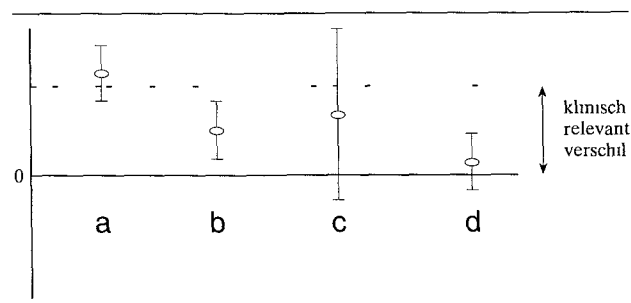
toetsuitslag	effect	geen effect	voorspellende waarde van een significante uitslag
<i>voorbeeld 1</i>			
significant	80	5	80/85 = 94%
niet significant	20	95	
totaal	100	100	
<i>voorbeeld 2</i>			
significant	80	5 000 000	80/5 000 080 = 0,002%
niet significant	20	95 000 000	
totaal	100	100 000 000	

terafkans bij een statistisch significant onderzoeksresultaat nog steeds minder dan 1 op 10 000 is. Dit impliceert tevens dat onderzoek naar extreem onwaarschijnlijke effecten zinloos is, aangezien een positief effect deze onwaarschijnlijkheid niet temetdoet. Dit geldt a fortiori voor onmogelijke effecten, waarbij de vooraf- en achterafkans per definitie nul zijn.

Onderzoek vooral zinvol bij twijfel De Bayesiaanse kijk doet begrijpen hoe dezelfde resultaten, uit een enkel onderzoek of uit een meta-analyse, bij weldenkende deskundigen tot andere interpretaties en conclusies kunnen leiden. Dit komt dan door de a-priori-verwachtingen die de verschillende experts met betrekking tot een bepaalde hypothese koesteren. Het wordt eveneens duidelijk waarom onderzoeksuitslagen geen enkel effect zullen hebben op degenen die reeds absoluut overtuigd zijn van de (on)juistheid van een bepaalde hypothese. Dit betekent dat onderzoek de meeste zin heeft wanneer er gerede twijfel bestaat: dan is de toename van zekerheid, en de daaraan gekoppelde kans op beleidsveranderingen, het grootst.

Klinische relevantie van onderzoeksbevindingen De vraag of er een effect is, als een zwart-witfenomeen, is – behalve bij zaken als homeopathie of paranormale verschijnselen – zelden voldoende, het is bij de vergelijking van twee geneesmiddelen zeer waarschijnlijk dat er wel enig verschil in effect zal zijn. De vraag is echter, en dit geldt evenzeer voor het effect van risicofactoren, of deze effecten klinisch relevant zijn. In dat geval willen we niet weten of het middel werkte, dan wel beter werkte, maar hoe groot dat verschil in effect was. Vervolgens kan afgewogen worden of het de moeite waard is het middel te gebruiken, waarbij ook bijwerkingen, logistieke en financiële overwegingen een rol zullen spelen. Enerzijds is het dus goed mogelijk dat een verschil tussen twee therapieën, statistisch significant of niet, ons in het geheel niet interesseert omdat het niet klinisch relevant is. Anderzijds kan een onderzoek met een grote statistische onzekerheidsmarge, waardoor de gevonden verschillen niet significant zijn, ons vanwege de suggestie van een sterk, klinisch relevant effect juist bijzonder interesseren.

Betrouwbaarheidsinterval Het is onder meer hierom dat het de voorkeur verdient niet te veel aandacht aan significanties en p-waarden te besteden, doch te kijken naar de grootte van het effect van de bestudeerde factor en het bijbehorende betrouwbaarheidsinterval (figuur 2). Dit interval is een maat voor de statistische onzekerheid, en kan gezien worden als het gebied van plausible uitkomsten. In figuur 2 wordt de relatie tussen klinische relevantie en statistische significantie nader toegelicht. Een verschil in effect is statistisch significant wanneer het betrouwbaarheidsinterval de neutrale waarde nul niet omvat (situatie weergegeven bij a en b). Zo'n significant verschil kan kleiner zijn dan hetgeen nog klinisch relevant wordt gevonden (b) en is dan van beperkte waarde. Wanneer een verschil relevant en significant is, leidt dit tot duidelijke conclusies, terwijl niet-significante verschillen (c en d) meer tot de verbeelding spreken wanneer ze klinisch relevant lijken (c). Dan dient men wel nader onderzoek te doen alvorens tot implementatie over te gaan.



FIGUUR 2 Het onderscheid tussen statistische significantie en klinische relevantie door toepassing van het betrouwbaarheidsinterval²³. De verticale lijnstukken vertegenwoordigen het 95%-betrouwbaarheidsinterval rondom de puntschatting van het verschil in effect tussen twee interventies: (a) het verschil is statistisch significant en klinisch relevant, (b) het verschil is statistisch significant, maar klinisch irrelevant, (c) het verschil is niet statistisch significant en de klinische relevantie is onduidelijk, (d) het verschil is niet statistisch significant en het is klinisch irrelevant.

DE ULTIEME WAARHEID

Een intrinsiek probleem bij het schrijven over 'ware' en 'onware' conclusies uit onderzoek is dat met het onderzoek dat beschreven wordt de waarheid alleen maar benaderd kan worden. Het is principieel onmogelijk empirisch aan te tonen dat de kans op colonkanker toeneemt met de duur van de colitis ulcerosa, of dat homeopathie onwerkzaam is. Het zou onjuist zijn het primaat aan de empirie te verlenen, in de zin dat hier absolute bewijskracht aan wordt verleend. Het zal altijd gaan om een combinatie van empirische bevindingen en theoretische overtuigingen, waarbij volledige zekerheid onbereikbaar is. Overigens hoeft dit inzicht niet te leiden tot pessimisme of nihilisme: de omgang met onzekerheden vormt de kern van de geneeskunde en voldoet prima in het dagelijks gebruik. Het is van groot belang dat men onderzoek verricht met zo min mogelijk methodologische tekortkomingen, teneinde de kans op een resultaat dat strookt met de waarheid te maximaliseren. Net zo min echter als een onderzoek met methodologische tekortkomingen onvermijdelijk leidt tot een onjuiste conclusie, leidt een methodologisch perfect onderzoek altijd tot een juist en relevant resultaat. De waarheid blijft ongreepbaar, steeds wenkend en steeds wijkend.

ABSTRACT

Roaming through methodology (conclusion) XXXIX The ultimate truth

- Studies that have methodological shortcomings may still lead to the right conclusion. Sometimes, the bias is small, or the shortcoming is not relevant for the issue being studied.
- It is also possible that a perfectly designed and executed study does not yield the 'right' answer. For example, chance circumstances might lead to the incomparability of groups.
- Bayes' theorem clarifies why different experts may come to opposite conclusions based on the same studies.
- The ultimate truth remains elusive.

LITERATUUR

- ¹ Schultz KF, Chalmers I, Hayes RJ, Altman DA. Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled clinical trials. *JAMA* 1995; 273:408-12.
- ² Lijmer JG, Mol BW, Heisterkamp S, Bossel GJ, Prins MH, Meulen JHP van der, et al. Empirical evidence of design related bias in studies of diagnostic tests. *JAMA* 1999;282:1061-6.
- ³ Linde K, Scholz M, Ramirez G, Clausius N, Melchart D, Jonas WB. Impact of study quality on outcome in placebo-controlled trials of homeopathy. *J Clin Epidemiol* 1999;52:631-66.
- ⁴ Greenstein AJ, Sachar DB, Smith H, Pucillo A, Papatestas AE, Kreef I, et al. Cancer in universal and left-sided ulcerative colitis: factors determining risk. *Gastroenterology* 1979;77:290-4.
- ⁵ Sackett DL, Whelan G. Cancer in ulcerative colitis: scientific requirements for the study of prognosis. *Gastroenterology* 1980;78:1632-5.
- ⁶ Tugwell PX. Hoe moeten medische tijdschriften worden gelezen? III. Het beoordelen van het klinische beloop en de prognose van een ziekte. *Ned Tijdschr Geneesk* 1983;127:2367-70.
- ⁷ Katzka I, Brody RS, Morris E, Katz S. Assessment of colorectal cancer risk in patients with ulcerative colitis: experience from a private practice. *Gastroenterology* 1983;85:22-9.
- ⁸ Ekblom A, Helmick C, Zack M, Adam H. Ulcerative colitis and colorectal cancer: A population based study. *N Engl J Med* 1990; 323:1228-33.
- ⁹ Jick H, Slone D, Westerholm B, Inman WHW, Vessey MP, Shapiro S, et al. Venous thromboembolic disease and ABO blood type: A cooperative study. *Lancet* 1969;1:539-42.
- ¹⁰ Lensing AW, Prandoni P, Brandjes D, Huisman PM, Vago M, Tomascella G, et al. Detection of deep-vein thrombosis by real-time B mode ultrasonography. *N Engl J Med* 1989;320:342-5.
- ¹¹ Huisman MV, Buller HR, Cate JW ten Vreeken J. Serial impedance plethysmography for suspected deep venous thrombosis in outpatients: The Amsterdam General Practitioner Study. *N Engl J Med* 1986;314:823-8.
- ¹² Koster T, Blann AD, Briet E, Vandenbroucke JP, Rosendaal FR. Role of clotting factor VIII in effect of von Willebrand factor on occurrence of deep vein thrombosis. *Lancet* 1995;345:152-5.
- ¹³ Allaart CF, Poort SR, Rosendaal FR, Reitsma PH, Bertina RM, Briet E. Increased risk of venous thrombosis in carriers of hereditary protein C deficiency defect. *Lancet* 1993;341:134-8.
- ¹⁴ Lenssen RP, Bertina RM, Ronde H de, Vandenbroucke JP, Rosendaal FR. Venous thrombotic risk in family members of unselected individuals with factor V Leiden. *Thromb Haemost* 2000;83:817-21.
- ¹⁵ Boven HH van, Vandenbroucke JP, Briet E, Rosendaal FR. Gene-gene and gene-environment interactions determine risk of thrombosis in families with inherited antithrombin deficiency. *Blood* 1999;94:2590-4.
- ¹⁶ Koes BW, Bouter LM, Mameren H van, Essers AH, Verstegen GM, Hofhuizen DM, et al. The effectiveness of manual therapy, physiotherapy, and treatment by the general practitioner for nonspecific back and neck complaints: A randomized clinical trial. *Spine* 1992; 17:28-35.
- ¹⁷ Koes BW, Bouter LM, Mameren H van, Essers AH, Verstegen GM, Hofhuizen DM, et al. Randomised clinical trial of manipulative therapy and physiotherapy for persistent back and neck complaints: results of one year follow up. *BMJ* 1992;304:601-5.
- ¹⁸ Kleijnen J, Knipschild P, Riet G ter. Clinical trials of homeopathy. *BMJ* 1991;302:316-23.
- ¹⁹ Linde K, Clausius N, Ramirez G, Melchart D, Eitel F, Hedges LV, et al. Are the clinical effects of homeopathy placebo effects? A meta-analysis of placebo-controlled trials. *Lancet* 1997;350:834-43.
- ²⁰ Oberg J. In: Sagan C. The Demon haunted world: science as a candle in the dark. New York: Random House, 1996. p. 187.
- ²¹ Wulff HR. Rational diagnosis and treatment: an introduction to clinical decision making. Oxford: Blackwell, 1976.
- ²² Tijssen JGP. Statistische methoden. In: Lubben J, Lang R de, redacteuren. *Klinisch geneesmiddelenonderzoek*. Utrecht: Bunge, 1987. p. 117-42.
- ²³ Berry G. Statistical significance and confidence intervals. *Br J Clin Pract* 1988;42:465-8.

Anvaard op 13 september 2001