'Het is wetenschappelijk bewezen'

Over significantie, relevantie en geloof

N.P. VAN DUIJN
A.A.M. HART

Van Duijn NP, Hart AAM. 'Het is wetenschappelijk bewezen'; over significantie, relevantie en geloof. Huisarts Wet 1994; 37(5): 176-80.

Samenvatting Het traject van onderzoeksresultaat naar toepassing in de praktijk telt niet alleen 'harde' cijfers, maar ook 'zachte' beoordelingsmomenten. De relevantie van de vraagstelling is geen feit, maar een oordeel, dat mede afhankelijk is van het gezichtspunt dat men inneemt. Ook een reeds bestaande overtuiging speelt een rol, soms zelfs zo sterk, dat er geen onderzoeksresultaat tegen bestand is. De p-waarde is wel een maat voor de zekerheid dát er een verschil is gevonden, maar zegt niets over de klinische relevantie van dat verschil; die moet afzonderlijk beoordeeld worden. De interpretatie van het gevonden verschil hangt vervolgens samen met de wijze waarop het verschil wordt weergegeven; bovendien moet er een uitspraak worden gedaan over de nauwkeurigheid van het gevonden verschil. Iets is dus nooit 'wetenschappelijk bewezen'; het is hooguit meer of minder aannemelijk gemaakt. Daarna kan men zijn opvatting herzien. Of niet.

Universiteit van Amsterdam, Meibergdreef 15, 1105 AZ Amsterdam.
Dr. N.P. van Duijn, huisarts-epidemioloog, Vakgroep Huisartsgeneeskunde;
Ir. A.A.M. Hart, statisticus, afdeling Klinische Epidemiologie en Biostatistiek, Academisch Medisch Centrum.
Correspondentie: Dr. N.P. van Duijn.

Inleiding

Numbers don't remember were they came from[1]

Wetenschappelijke bewijsvoering draait vaak om de vraag of er een verschil is tussen bijvoorbeeld twee aandoeningen of twee behandelingen, en of dat verschil 'statistisch significant' is. Daarbij worden dan vaak exacte p-waarden vermeld, zoals p=0,003. Dit suggereert dat het verschil 'dus' bewezen is.

'Dus' omvat echter meer dan alleen statistische significantie. 'Dus' heeft ook betrekking op de kwaliteit van de wijze waarop de gegevens verkregen zijn, de wetenschappelijke gegevens zelf, de cijfermatige bewerkingen om de gegevens weer te geven, de nauwkeurigheid van de uitkomst, en de interpretatie van dit geheel binnen dit specifieke onderzoek. En als het verschil op grond van die interpretatie geaccepteerd wordt als 'een echt verschil tussen de groepen patiënten in dit onderzoek', moeten nóg twee vragen beantwoord worden: is de grootte van het gevonden verschil relevant voor de praktijk, en mag dit resultaat toegepast worden op vergelijkbare patiënten in een andere praktijk? Deze hele redenering van onderzoeksresultaat naar toepassing in de praktijk bevat dus, naast harde gegevens, een aantal 'zachte' beoordelingsmomenten waarover de meningen kunnen verschillen.

In feite zijn er nog meer 'zachte' elementen. De meest basale vraag is natuurlijk of de lezer de vraagstelling wel relevant vindt. Daarnaast heeft de lezer meestal al een – min of meer expliciet – oordeel over het betreffende onderwerp. Als hij tevoren al sterk gelooft in het bestaan van een verschil, zal er naar verhouding weinig voor nodig zijn om hem in dat geloof te bevestigen. Is hij er daarentegen van overtuigd dat er geen verschil is, dan kan alleen een krachtige bewijsvoering invloed hebben op zijn ongeloof.

In dit artikel worden zes elementen van bewijsvoering en toepassing besproken. Aan de hand van een voorbeeld uit de literatuur over de nadelen van koffie (*kader*) wordt duidelijk gemaakt wanneer de lezer zelf een keuze kan maken en hoe een dergelijke redenering kan leiden tot verschillende, genuanceerde conclusies. Het gaat om:

- de relevantie van de vraagstelling;
- de overtuiging op voorhand;
- de statistische significantie;
- de klinisch relevante grootte van het verschil;
- de weergave van de grootte van het verschil;
- de nauwkeurigheid van de uitspraak over het verschil.

Relevantie van de vraagstelling

Leidt koffiedrinken tot een verhoogd cholesterolgehalte?[2-5]

Dit lijkt een relevante vraagstelling, want een verhoogd cholesterolgehalte is belangrijk en koffie wordt veel gedronken. Sommige lezers zullen deze vraag zeer relevant vinden uit interesse voor de etiologie van hypercholesterolemie, terwijl anderen de vraagstelling van minder groot belang zullen achten, omdat de morbiditeit en de mortaliteit niet duidelijk verhoogd zijn volgens de literatuur.[3,6-8] Sommige lezers gaan misschien nog verder en vinden de vraagstelling pas relevant als het advies geen koffie te drinken een belangrijke bijdrage levert aan de preventie van coronaire aandoeningen.[9]

De relevantie van de vraagstelling is dus een kwestie van beoordeling vanuit een bepaald gezichtspunt: er is geen absoluut gelijk of ongelijk.

Overtuiging op voorhand

Veel lezers zullen vinden dat koffiedrinken op zichzelf niet slecht zal zijn voor hart- en bloedvaten, maar dat overdaad schaadt; anderen zijn voorzichtiger en adviseren voor de zekerheid coffeïnevrije koffie, maar niemand zal menen dat veel koffiedrinken een zeer gezonde gewoonte is. Lezers kunnen zo sterk verschillen in hun overtuiging op voorhand, dat geen enkele bewijsvoering invloed op hun

standpunt kan uitoefenen.[10,11] Op die wijze is onlangs een doortimmerd huisartsgeneeskundig betoog over terughoudendheid met ACE-remmers en Ca-antagonisten bij hypertensie zonder enige gêne aan de kant geschoven.[12,13] Een voornaam argument van de commentator luidde: '... één ding is zeker, vooruitgang is niet te stuiten'. Een dergelijke extreme overtuiging op voorhand betekent een regelrechte bedreiging voor elk wetenschappelijk debat.

The effect on serum cholesterol levels of coffee brewed by filtering or boiling
Bak AAA, Grobbee DE. N Engl J Med 1989; 321: 1432-7

Inleiding De literatuur duidt op een verband tussen koffie drinken en een verhoogd cholesterolgehalte. Dit geldt vooral voor 'gekookte' koffie ('boerenkoffie': kokend water op de koffie, na 10 minuten afschenken zonder filter). Om hierover meer zekerheid te krijgen, is een vergelijkend onderzoek uitgevoerd.
Methode 107 gezonde jong-volwassenen werden toevalsgewijs verdeeld over drie groepen. Eén groep dronk 4-6 kopjes boerenkoffie per dag, een tweede groep dronk 4-6 kopjes filterkoffie en een derde groep dronk geen koffie. Voor dit onderzoek zijn speciale kopjes van 140 ml uitgereikt. Na randomisatie bleken de drie groepen goed vergelijkbaar te zijn. Het voorgeschreven koffiegebruik is streng gecontroleerd door middel van dagboeken, tellen van koffieverpakkingen en onaangekondigde huisbezoeken waarbij speekselmonsters genomen werden ter bepaling van het coffeïnegehalte. Tevoren en na 3, 6 en 9 weken is het totaal, HDL- en LDL-cholesterolgehalte bepaald.
Resultaten In de groep die boerenkoffie gebruikte, was het totaal cholesterolgehalte na 9 weken met 0,48 mmol/l (95%-betrouwbaarheidsinterval 0,13-0,83) gestegen. Het LDL-cholesterol was 0,39 mmol/l (95%-betrouwbaarheidsinterval –0,04-0,82) gestegen. Bij gebruikers van filterkoffie en bij personen die geen koffie dronken, was geen cholesterolstijging opgetreden.
Conclusie Boerenkoffie leidt tot een gemiddelde stijging van het cholesterolgehalte van 10 procent, vergeleken met het drinken van filterkoffie of geen koffie.

Coffee, caffeine, and cardiovascular disease
Grobbee DE, Rimm EB, Giovannucci E, et al. N Engl J Med 1990; 323: 1026-32.

Inleiding De literatuur duidt op enig risico van koffie voor cardiovasculaire aandoeningen, hoewel veel resultaten tegenstrijdig zijn.
Methode In een cohort van 45.589 mannen van 40-75 jaar zonder cardiovasculaire voorgeschiedenis is prospectief het verband onderzocht tussen koffie-consumptie en myocardinfarct, bypass-operatie, angioplastiek of CVA na twee jaar.
Resultaten Het relatief risico van vier of meer kopjes koffie dagelijks vergeleken met geen koffie drinken voor het optreden van cardiovasculaire aandoeningen is 0,81 (95%-betrouwbaarheidsinterval 0,56-1,18). Meer of minder koffie drinken leidt niet tot een hoger of lager risico.
Conclusie Deze resultaten geven geen steun aan de veronderstelling dat koffie leidt tot cardiovasculaire aandoeningen.

Statistische significantie

Een statistische toets is te vergelijken met een diagnostische test op bijvoorbeeld een verschil tussen twee groepen ('ziek' en 'niet ziek').[4] De beste vergelijking levert een diagnostische test waarvan men de specificiteit kan bepalen door zelf het afkappunt tussen 'ziek' en 'niet ziek' te kiezen. Deze statistische diagnostiek wordt als positief beschouwd – dat wil zeggen: statistisch significant – als de p-waarde heel klein is, bijvoorbeeld <0,05 of zelfs <0,01. Net als bij een diagnostische test komen bij de statistische diagnostiek ook fout-positieve en fout-negatieve uitslagen voor (*tabel 1*).

Uit de diagnostiek is bekend dat we weinig hebben aan een test die niet gevoelig genoeg is. In de statistische diagnostiek geldt hetzelfde. Een statistische test op een verschil is ongevoelig (lage sensitiviteit) bij een te kleine onderzoeksgroep. Bij een klein onderzoek wordt dus gemakkelijk een werkelijk verschil gemist. Maar áls in een kleine onderzoeksgroep een statistisch significant verschil gevonden wordt, is dat voldoende. Men kan geluk hebben, zoals men ook geluk kan hebben met nierpalpatie (een ongevoelige test voor een niervergroting): als men de nier goed voelt, heeft dat betekenis, maar voelt men niets bijzonders, dan zegt dat niets.

Een p-waarde van <0,003 is dus slechts één element in een wetenschappelijke bewijsvoering – soms een belangrijk, soms een minder belangrijk element.[1,14]

Klinisch relevante grootte van het verschil

Het maakt nogal verschil of een etiologische factor als koffie het cholesterolgehalte gemiddeld 3 mmol/l doet stijgen of slechts 0,2 mmol/l. Wat men hoog vindt, is een kwestie van beoordeling op basis van kennis van het onderwerp, een kwestie van relevantie dus. Het klinisch relevante verschil kan per vraagstelling en per discipline verschillen. In de curatieve zorg kan de klinische relevantie van een verschil zelfs per patiënt verschillen en wordt zij dus in feite mede door de patiënt bepaald.

Wanneer er binnen een discipline uiteenlopende opvattingen over de klinische relevantie van verschillen bestaan, kan dit leiden tot heftige discussies, zoals destijds over de screening op hypertensie. Veel stellingen in NHG-standaarden zijn het resultaat van debat en consensus over klinisch relevante verschillen. Daarom is het maken van een NHG-standaard zo leuk.

Weergave van de grootte van het verschil

Men kan een verschil alleen beoordelen op zijn relevantie, als de grootte van het verschil inzichtelijk is weergegeven. Het relatief risico kan een bruikbare maat zijn bij dichotome resultaten: eenvoudige ja/nee-gegevens. Het relatief risico wordt vooral gebruikt in etiologisch onderzoek: vandaar de naam. Bij therapeutisch onderzoek zou men de term 'relatief effect' kunnen gebruiken.

In *tabel 2* wordt een voorbeeld gegeven. Het risico op hart- en vaataandoeningen voor koffiedrinkers is 7,2 per 1000 personen per jaar; voor niet-koffiedrinkers is het 8,9 per 1000 per jaar. Het relatief risico is de breuk van deze twee risico's: 0,81. De koffiedrinkers hadden in dit onderzoek dus 19 procent minder risico op hart- en vaataandoeningen. Maar die 19 procent is relatief: een beetje van een beetje is een klein beetje. Eigenlijk gaat het om twee patiënten minder op 1000 mensen. Dat is de absolute risicovermindering. Voorstanders van screening hanteren de relatieve risicovermindering, tegenstanders de absolute risicovermindering. Beiden hebben gelijk, maar één krijgt het.

In *tabel 2* zijn de resultaten op verschillende manieren weergegeven. Welke weergave men ook kiest, de oorspronkelijke gegevens behoren zodanig beschreven te worden, dat de lezer de wijze van weergave kan controleren. Want het is de lezer die uiteindelijk moet (kunnen) kiezen.

Nauwkeurigheid van de uitspraak over een verschil

Een conclusie over een verschil is mede afhankelijk van de nauwkeurigheid van de gegevens, en die nauwkeurigheid wordt voor een deel bepaald door het toeval. De mate waarin dat gebeurt, kan worden weergegeven met het betrouwbaarheidsinterval (in de praktijk meestal het 95%-betrouwbaarheidsinterval).

Deze maat voor nauwkeurigheid is niet nieuw – *Neyman & Pearson* introduceerden het 95%-betrouwbaarheidsinterval al in 1930[15] –, maar hij wordt pas de laatste jaren op ruime schaal toegepast. De definitie van het 95%-betrouwbaarheidsinter-

Tabel 1 *Significantietoets opgevat als diagnostische test, uitgevoerd op data van een vergelijkend onderzoek en afgezet tegen de 'werkelijkheid'*

		'Werkelijkheid'	
		verschil	geen verschil
Statistische toets	p<0.05	terecht-positief a	fout-positief b
	p ≥0.05	fout-negatief c	terecht-negatief d

Sensitiviteit: $\frac{a}{a+c}$. Specificiteit: $\frac{d}{b+d}$.

In deze analogie is de specificiteit van de statistische toets het percentage niet-significanties bij geen verschil tussen de twee groepen: d/(b+d). Het significantieniveau, bijvoorbeeld <0,05, kan dus ook gezien worden als een specificiteit van de statistische toets van >95 procent. De sensitiviteit is dan het percentage significanties bij een werkelijk verschil tussen de twee groepen: a/(a+c).
De sensitiviteit is afhankelijk van de grootte van de groepen patiënten. Hoe groter de aantallen onderzochte patiënten, des te sensitiever de statistische toets is om een reëel verschil tussen de groepen patiënten te ontdekken. Hier wordt de term *onderscheidend vermogen* ('power') gebruikt. Het onderscheidend vermogen is dan 1 – sensitiviteit, of c/(a+c).
Voor de fijnproevers tenslotte: de prevalentie in deze analogie komt overeen met de overtuiging op voorhand, het professioneel geloof van de arts.

Tabel 2 *Het relatief risico van koffiedrinken op cardiovasculaire morbiditeit (myocardinfarct, ingreep coronairvaten, CVA)*

	Cardiovasculair incident		
	ja	nee	totaal
≥4 kopjes koffie per dag	36	4.950	4.986
geen koffiedrinker	116	12.940	13.056

Relatief risico: $\frac{36/4.986}{116/13.056} = \frac{7,2/1000}{8,9/1000}$ =	0,81
95%-betrouwbaarheidsinterval	0,56-1,18
p-waarde	ca. 0,25
procentuele risicovermindering	19 procent
absolute risicovermindering	1,7 per 1000 personen per jaar
aantal mensen dat 4 kopjes koffie moet gaan drinken om 1 cardiovasculair incident te voorkomen	588

val rond een relatief risico is: 'een interval rond het waargenomen relatieve risico van mogelijke waarden die, gegeven het 95%-criterium, geacht mogen worden consistent te zijn met de waarnemingen op basis van de onzekerheid van het resultaat door toevalsvariabiliteit'.[4] Dit betekent dat het betrouwbaarheidsinterval alleen iets zegt over de nauwkeurigheid van déze uitkomst, gegeven de toevalsvariatie in déze studie bij déze aantallen patiënten. Er wordt dus geen uitspraak gedaan over de in feite onbekende 'werkelijke' uitslag. Als grote aantallen patiënten onderzocht worden, is de rol van het toeval kleiner dan in een onderzoek bij 10 patiënten. Bij kleine studies is het 95%-betrouwbaarheidsinterval daardoor groot, bij grote studies klein.

Voor de berekening van het betrouwbaarheidsinterval zijn formules beschikbaar, die zijn ontleend aan de kanstheorie.[16] In het voorbeeld van koffiedrinken is het 95%-betrouwbaarheidsinterval 0,56–1,18 rond het gevonden relatief risico van 0,81 (*tabel 2*). Het is dus niet erg aannemelijk – op basis van déze studie – dat het werkelijke relatieve risico toevallig kleiner is dan 0,56 of groter dan 1,18.

Beschouwing

Wie kennis wil generaliseren naar de eigen patiënten, heeft niet alleen te maken met onderzoeksresultaten, maar ook met overtuigingen op voorhand en beoordelingsmomenten. De redenering van onderzoeksresultaat naar toepassing zou er voor het voorbeeld in *tabel 2* als volgt uit kunnen zien.

- *Zijn de vraagstellingen 'Verhoogt koffiedrinken het cholesterolgehalte?' en 'Verhoogt koffiedrinken de cardiovasculaire morbiditeit?' relevant?*

Over de relevantie van de eerste vraagstelling kunnen verschillen in opvatting bestaan. Een lezer met een theoretische belangstelling voor de etiologie van hypercholesterolemie zal deze vraagstelling relevant vinden. Wie meer praktisch georiënteerd is, zal de vraag niet relevant vinden en het artikel dus ongelezen laten.

De tweede vraag zal door sommigen zonder meer bevestigend worden beantwoord. Anderen zullen alleen de omgekeerde vraag relevant vinden: 'Verlaagt het stoppen met koffiedrinken de morbiditeit?' Beleidmakers kunnen nog verder gaan door alleen artikelen te lezen met de vraagstelling 'Verlaagt een koffiestop-advies de morbiditeit?'.

- *Had u tevoren gedacht dat koffiedrinken een risicofactor was?*

Dit moet ieder voor zich uitmaken. Het is wel belangrijk zich zo'n 'vooraf-geloof' te realiseren. De bijtende toon van sommige discussies over de interpretatie van onderzoeksresultaten is hiermee te verklaren.[9 13] Het komt zelfs voor dat kritische commentaren geheel bestaan uit het etaleren van een vooraf-geloof.[13] Frisse kritiek is dan ontaard in een vermanende preek.

- *Is er een significant verschil tussen koffiedrinkers en niet-koffiedrinkers?*

Wel naar cholesterolgehalte, niet naar morbiditeit.[2 3 5 6]

- *Is het tussen koffiedrinkers en niet-koffiedrinkers gevonden verschil een relevant verschil?*

Een verschil van 0,5 mmol/l zal door sommigen als te klein beoordeeld worden; zij zullen tegen patiënten zeggen dat koffie niet goed of slecht is voor hart- en bloedvaten. Anderen zullen het verschil in cholesterolgehalte groot genoeg vinden om een interventiestudie te overwegen. Wellicht zullen sommigen het verschil groot genoeg vinden om koffie voortaan te ontraden.

- *Hoe veel draagt koffiedrinken precies bij aan de morbiditeit?*

Hoewel het cholesterolgehalte van koffiedrinkers gemiddeld hoger is dan dat van niet-koffiedrinkers, is er volgens *tabel 2* geen sprake van een hoger relatief risico met betrekking tot de morbiditeit. Integendeel, het gevonden relatief risico is juist lager. Koffieverkopers zouden op grond van dit onderzoek in hun reclame kunnen stellen, dat 'veel koffiedrinken 19 procent gezonder is'. Anderen spreken liever van 2 cardiovasculaire incidenten minder per 1000 patiënten per jaar. Een derde mogelijkheid is te stellen dat ongeveer 600 mensen (veel) meer koffie moeten gaan drinken om één cardiovasculair incident te voorkomen (*tabel 2*).

- *Hoe nauwkeurig is het resultaat van deze studie?*

Als we de gekozen onderzoeksopzet accepteren, is het resultaat redelijk nauwkeurig te noemen. Als iedereen koffie gaat drinken, kan er per praktijk misschien één cardiovasculair incident per jaar minder optreden. Twee minder, of misschien zelfs één meer kan ook.

Alles bijeen is het veronderstelde risico van koffiedrinken op cardiovasculaire morbiditeit en mortaliteit niet aannemelijk gemaakt door dit onderzoek.

Conclusie

Iets is nooit 'wetenschappelijk bewezen'. Opvattingen vooraf, nieuwe feiten en de interpretatie van die feiten leiden tot een conclusie, die noch zeker, noch objectief is. Op grond daarvan kan men zijn opvattingen wijzigen. Of niet.

Dankbetuiging

Met dank aan Prof.dr. J.G.P. Tijssen, klinisch epidemioloog, voor zijn ideeën en het verhelderende commentaar.

Literatuur

1 Andersen B. Methodological errors in medical research. Oxford: Blackwell, 1990.
2 Bak AAA, Grobbee DE. The effect on serum cholesterol levels of coffee brewed by filtering or boiling. N Engl J Med 1989; 321: 1432-7.
3 Grobbee DE, Rimm EB, Giovannucci E, et al. Coffee, caffeine, and cardiovascular disease. N Engl J Med 1990; 323: 1026-32.
4 Tijssen JGP. Methodologie van klinisch onderzoek in de cardiologie. Utrecht: Bunge, 1992.
5 Fried RE, Levine DM, Kwiterovich PO, et al. The effect of filtered-coffee consumption on plasma lipid levels. JAMA 1992; 267: 811-5.
6 Grobbee DE, Rimm EB, Colditz G, et al. Coffee, caffeine, and cardiovascular disease. N Engl J Med 1991; 324: 992.
7 Van Dusseldorp M, Katan MB. Het effect van koffie op het serumcholesterolgehalte. Ned Tijdschr Geneeskd 1990; 134: 2325-7.

8 Verhoeff FH, Millar JM. Does caffeine contribute to postoperative morbidity? Lancet 1990; 336: 632.
9 Meijler FL. Het effect van koffie op het serumcholesterolgehalte. Ned Tijdschr Geneeskd 1991; 135: 391.
10 Vandenbroucke JP. Use and misuse of the biomedical literature. In: Proceedings of future trends of biomedical documentary information, March 8, 1991. Leiden: University of Leiden, Boerhaave Committee, 1991: 3.
11 Anonymous. Subjectivity in data analysis. Lancet 1991; 337: 401-2.
12 Grundmeijer HGLM. Waarom toch diuretica en bètablokkers als eerste keus bij hypertensiva? Hart Bull 1993; 24: 196-9.
13 Man in 't Veld AJ. Commentaar. Hart Bull 1993; 24: 199-201.
14 Andersen B, Holm P. Problems with p; significance testing in medical research. Basel: Hoffmann la Roche, 1984.
15 Good IJ. Good thinking; the foundations of probability and its applications. Minneapolis: University of Minnesota Press, 1983.
16 Gardner MJ, Altman DG. Statistics with confidence. Confidence intervals and statistical guidelines. Br Med J 1989. Met het bijhorende software pakket CIA. ■

Abstract

Van Duijn NP, Hart AAM. 'Scientifically proved'. On significance, relevance and belief. Huisarts Wet 1994; 37(5): 176-80.

Scientific inference and the decision to apply scientific results to patient care contain various elements that require a professional judgement. With the literature on the association between coffee and cardiovascular disease as an example, six of these elements are discussed: relevance of the question, prior belief, statistical significance, clinical relevant size of a difference, relative or absolute difference, and precisement of the difference as found. Statistical significance is discussed analogous to diagnostic tests. The p-value designates the certainty that an actual difference was found, but has no bearing on the size of the difference. Imprecisement by chance is denoted by a 95% confidence interval. The six elements combined lead to generalisation of the results. It is stressed that scientific inference is based on many normative judgements. A conclusion has no absolute scientifical validity. Instead, professionals assess the validity of a conclusion and consequently decide whether or not to change their prior belief.

Correspondence Dr. N.P. van Duijn, Department of General Practice, University of Amsterdam, Meibergdreef 15, 1105 AZ Amsterdam, The Netherlands.