Stop met het onkritische gebruik van nulhypothesen

de Boer, Michiel; van Grootel, Leonie; Bouter, Lex

published in
Huisarts en Wetenschap
2018

DOI (link to publisher)
10.1007/s12445-018-0255-4

document version
Publisher's PDF, also known as Version of record

Stop met het onkritische gebruik van nulhypothesen

Michiel de Boer, Leonie van Grootel, Lex Bouter

Conclusies op basis van medisch-wetenschappelijk onderzoek zijn dikwijls onjuist. Een belangrijke oorzaak is dat onderzoekers conclusies vaak trekken op basis van nulhypothesesignificantietoetsen (NHST). Veel artsen die hun praktijk afstemmen op wetenschappelijk onderzoek beseffen echter niet dat de zeggingskracht van de gebruikte p-waarden vaak beperkt is. Wij pleiten ervoor om te stoppen met het onkritische gebruik van NHST en het dichotome denken dat hieraan ten grondslag ligt. Onderzoekers zouden eigenlijk meer oog moeten hebben voor de grootte van verschillen en de (on)waarschijnlijkheid daarvan.

ONS ONWANKELBARE GELOOF IN P-WAARDEN

Statistiek wordt door menig arts gezien als een noodzakelijk kwaad. In de geneeskundeopleiding is er relatief weinig aandacht voor statistiek vanwege de grote competitie met medisch-inhoudelijke vakken en het trainen van essentiële vaardigheden voor de praktijk. Het is dan ook niet verwonderlijk dat vele artsen al heel tevreden zijn wanneer ze een aantal basisprincipes in de praktijk kunnen toepassen. Dan is het vooral belangrijk om te kunnen concluderen dat bijvoorbeeld de effectiviteit van een behandeling bewezen is. Voor dat laatste is er een handige vuistregel: wanneer de p-waarde kleiner is dan 0,05 ($p < 0,05$), is aangetoond dat de behandeling inderdaad werkt.

Hoe waarschijnlijk is een bepaald verschil is een betere benadering dan te constateren of er wel of geen verschil is. Foto: iStock

DE KERN

- De p-waarde uit een statistische toets geeft ons niet de kans dat de nulhypothese waar of onwaar is, terwijl we daar wel een conclusie over trekken.
- De kans dat we terecht concluderen dat een bevinding 'significant' is, is vergelijkbaar met een voorspellende waarde van een positieve testuitslag: zonder informatie over de prevalentie zijn ze geen van beide uit te rekenen.
- Het gebruik van nulhypothesesignificantietoetsen heeft ertoe geleid dat een deel van de conclusies uit onderzoek onjuist is.
- Meer verantwoorde manieren van conclusies trekken vragen dat we onze dichotome manier van denken (wel of geen verschil) veranderen in een meer continue manier van denken (hoe waarschijnlijk is een bepaald verschil, hoe groot achten we dat verschil ongeveer en in welke mate zou het klinisch relevant kunnen zijn).

DE BEDRIEGLIJKHEID VAN INVERSE GEVOLGTREKKINGEN

Wat velen waarschijnlijk niet meer weten is wat die p-waarde, de nulhypothese en de alternatieve hypothese precies betekenen.[1] We gebruiken als voorbeeld het onderzoek van Bos en Buis uit het julinummer (2017) van dit tijdschrift, die automatische bloeddrukmeting gedurende 30 minuten (BD30) en conventionele spreekkamermetingen met elkaar hebben vergeleken.[2] Daarnaast hebben ze onderzocht of het voor de grootte van het verschil tussen beide methoden uitmaakte tot welke subgroep de patiënten behoorden (leeftijd, geslacht, diabetes mellitus en hart- en vaatziekten). De alternatieve hypothesen in dit voorbeeld wijzen erop dat deze verschillen bestaan. De nulhypothese betreft het ontbreken van de verschillen. Een van de bevindingen die de onderzoekers rapporteerden was dat het verschil tussen BD30 en spreekkamermeting aanzienlijk groter was bij patiënten ≥ 70, dan in de groep jongere patiënten. Voor de diastolische bloeddruk rapporteerden de auteurs een gemiddeld verschil van 6,2 mmHg ($p < 0{,}001$). De betreffende p-waarde betekent dan dat als de nulhypothese juist is en we dit onderzoek oneindig vaak zouden herhalen, minder dan 0,1% daarvan 6,2 mmHg of een groter verschil zou laten zien.

De meeste huisartsen die over dit onderzoek lezen, zullen op basis ervan concluderen dat het verschil tussen beide methoden van bloeddrukmeting dus groter is bij patiënten ≥ 70. De (impliciete) redenering die we hierbij volgen is dat als de nulhypothese waar zou zijn, de kans op het gevonden resultaat of een extremer resultaat zo klein is, dat we de nulhypothese moeten verwerpen. Precies op dit punt gaan we de mist in. We trekken een conclusie over de nulhypothese (verschillen zijn niet afhankelijk van leeftijd) op basis van een voorwaardelijke kans op de gevonden onderzoeksgegevens of extremer (P(gevonden verschil of groter|H0)), terwijl we dus eigenlijk de kans dat de nulhypothese waar is zouden willen bepalen.

Methodologen en statistici zijn al sinds nulhypothesesignificantietoetsen (NHST) bestaan bekend met dit probleem en het fenomeen heeft zelfs een officiële naam: de 'bedrieglijkheid van inverse gevolgtrekkingen'.[3,4] Dit besef is echter nog niet overal in de medische wetenschap doorgedrongen.

DE CONSEQUENTIES

Misschien vraagt u zich af waarom we hier zo'n probleem van maken. Wanneer de resultaten van het onderzoek (of een extremere uitkomst) onwaarschijnlijk zijn als de nulhypothese juist is, dan is het toch logisch dat het onwaarschijnlijk is dat de nulhypothese klopt? En dus ook dat er in werkelijkheid wel een invloed van leeftijd is, ofwel dat ons vermoeden, de alternatieve hypothese, klopt? In de praktijk blijkt echter dat dit niet altijd opgaat. Toegepast op ons voorbeeld komt dat doordat de zeggingskracht van de p-waarde voor de conclusie mede afhangt van de voorafkans dat er werkelijk een verschil is wanneer we beide methoden van bloeddrukmeting vergelijken tussen patiënten ≥ 70 of jonger. Dit is te vergelijken met de voorspellende waarde van een positieve testuitslag in de diagnostiek, die sterk afhangt van de voorafkans op de ziekte (de prevalentie).[5-7]

Neem het onderzoek van Schouten en Van de Putte in het afgelopen januarinummer (2018) van dit tijdschrift.[8] Zij onderzochten de validiteit van SPUTOVAMO-R2, een checklist voor kindermishandeling. Ze hebben onder andere vergeleken met een melding bij Veilig Thuis. In [**tabel 1**] staan in kruistabel A de gegevens zoals we die op basis van het artikel kunnen reconstrueren. De sensitiviteit is laag en de voorspellende waarde van een positieve testuitslag (VW+) ook. Wanneer de voorafkans (prevalentie) op kindermishandeling hoger is, zal de VW+ toenemen, ook al blijven sensitiviteit en specificiteit gelijk [**tabel 1**, kruistabel B]. Hetzelfde gebeurt met conclusies op basis van NHST. De power en betrouwbaarheid van de toets (1-α) zijn vergelijkbaar met respectievelijk de sensitiviteit en specificiteit. De kans dat we bij een significant resultaat terecht zullen concluderen dat er een verschil is (in ons eerste voorbeeld tussen patiënten ≥ 70 of jonger) zal toenemen wanneer de voorafkans hierop groter is, oftewel wanneer we het vooraf waarschijnlijker achten dat dit verschil er echt is [**tabel 1**, kruistabellen C en D]. Wanneer vooraf de kans klein geacht wordt (kruistabel C), dan is de VW+ rond de 0,5 – gelijk aan het opgooien van een muntje. Meestal hebben we bij onderzoek geen goed idee over wat de voorafkans op een werkelijk verschil is, en weten we in veel gevallen daarom niet wat de

> Velen weten waarschijnlijk niet meer wat die p-waarde, de nulhypothese en de alternatieve hypothese precies betekenen

Tabel 1

Een illustratie van de overeenkomst tussen de voorspellende waarde van een positieve diagnostische testuitslag en de zeggingskracht van een p-waarde

A	**VT+**	**VT-**		**B**	**VT+**	**VT-**	
Checklist+	9	99	VW+ = 0,01	Checklist+	27	97	VW+ = 0,18
Checklist-	478 Sens = 0,02 Prev = 0,01	50085 Spec = 0,998	VW- = 0,99	Checklist-	1434 Sens = 0,02 Prev = 0,03	49140 Spec = 1,0	VW- = 0,97
C	**Voorafkans echt verschil**	**Voorafkans geen verschil**		**D**	**Voorafkans echt verschil**	**Voorafkans geen verschil**	
Significant	0,05	0,045	VW+ = 0,53	Significant	0,25	0,025	VW+ = 0,91
Niet-significant	0,05 Power = 0,5* Voorafkans = 0,1	0,905 Betrouwbaarheid = 0,95	VW- = 0,95	Niet-significant	0,25 Power = 0,5 Voorafkans = 0,5	0,475 Betrouwbaarheid = 0,95	VW- = 0,66

De kruistabellen A en B laten zien dat de positief voorspellende waarde van de checklist voor het opsporen voor kindermishandeling toeneemt van 0,01 naar 0,18 wanneer de prevalentie (voorafkans) op kindermishandeling zou toenemen van 0,01 naar 0,03, uitgaande van gelijke sensitiviteit en specificiteit van de checklist. De kruistabellen C en D laten analoog daaraan zien dat de kans dat een significante toetsuitslag een werkelijk verschil aangeeft ook toeneemt naarmate de voorafkans daarop stijgt, uitgaande van gelijkblijvende power en betrouwbaarheid (1-α) van de toets.

VT = Veilig Thuis; Sens= sensitiviteit; Prev = prevalentie; Spec = specificiteit; VW+ = positief voorspellende waarde; VW- = negatief voorspellende waarde
* De power is hier op 0,5 gezet, aangezien de power voor het toetsen van subgroepverschillen vaak (fors) lager ligt dan de gebruikelijke 0,8 voor hoofdeffecten.

zeggingskracht van een p-waarde is. Als we op basis van die p-waarde dan wel een uitspraak doen over het verschil tussen beide bloeddrukmetingen, hebben we dus geen idee of die uitspraak klopt.

VERANTWOORDE CONCLUSIES TREKKEN OVER ONDERZOEKSGEGEVENS

Op basis van het bovenstaande blijkt dat we decennialang massaal een methode hebben toegepast die ertoe leidt dat een deel van de conclusies over onderzoek niet juist is. Hoewel de tekortkomingen van NHST al vaak naar voren zijn gebracht,[3-7,9-11] is er in de praktijk niet veel veranderd.[12] Er is nog weinig wetenschappelijk inzicht in de reden daarvan. Een van de vermoedelijke oorzaken is dat wetenschappers niet goed

We hebben decennialang massaal een methode toegepast die ertoe leidt dat een deel van de conclusies over onderzoek niet juist is

weten welke alternatieve methoden er zijn en ook de noodzaak niet voelen om zich hierin te verdiepen. Alternatieven zijn echter voorhanden en veel winst kan al gemaakt worden zonder dat daarvoor veel extra kennis of vaardigheden nodig zijn.

Onzes inziens is de grootste winst te bereiken wanneer we onze dichotome manier van denken (wel een verschil of geen verschil) veranderen in een meer continue manier van denken (hoe waarschijnlijk is een bepaald verschil, hoe groot achten we dat verschil ongeveer en in welke mate zou het klinisch relevant kunnen zijn). Een belangrijke stok achter de deur voor onderzoekers hierbij is het beleid van een aantal tijdschriften om het woord 'significant' te vermijden.[13] Dit vereist meer aandacht voor en interpretatie van beschrijvende gegevens.[14] Aangevuld met betrouwbaarheidsintervallen geeft dit een indicatie van de precisie van de verschillen of effecten, mits er niet getoetst wordt op basis van de intervallen. In ons voorbeeldartikel van Bos en Buis wordt dit alles al deels gedaan door de puntschattingen en spreiding van de boven- en onderdruk van zowel de patiënten ≥ 70 jaar als die van < 70 jaar te geven. Ook benoemen de auteurs in de interpretatie van deze subgroepvergelijking de grootte van de verschillen. In het originele artikel in *Family Practice* tabelleren de auteurs ook de andere subgroepvergelijkingen en geven ze aan dat ze geen verschillen hebben gevonden ($p > 0,15$).[15] Hier komen ze in de beschouwing of conclusie van het artikel helaas niet op terug. Dat maakt een betere interpretatie van de bevindingen niet mogelijk. De kruistabellen C en D van [**tabel 1**] laten immers zien dat afhankelijk van de waarschijnlijkheid van echte verschillen tussen de subgroepen, de kans op werkelijke verschillen bij significantie (kruistabel C) of de kans op het ontbreken daarvan (kruistabel D) behoorlijk laag kan worden.

Ten slotte willen we niet onvermeld laten dat er ook andere statistische benaderingen zijn, waarvan bayesiaanse methoden het bekendst zijn. Bij bayesiaanse statistiek wordt de a priori verwachting geïncorporeerd in het statistische model. Grootschaliger gebruik van bayesiaanse methoden is onlangs

mogelijk geworden dankzij gebruiksvriendelijke software (zie onder andere: https://jasp-stats.org). Een ander recentelijk beschreven alternatief voor NHST is de a priori inferentiemethode.[11] Deze lijkt op de welbekende *sample size*-berekening, maar heeft het voordeel dat achteraf geen toetsing meer nodig is. Een uitgebreidere inleiding in deze methoden is elders te vinden.[11,16] ■

LITERATUUR

1. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 2016;31:337-50. DOI:10.1007/s10654-016-0149-3.
2. Bos MJ, Buis S. Dertig-minutenbloeddrukmeting op de praktijk. Huisarts Wet 2017;60:320-2.
3. Fisher R. Statistical methods and scientific induction. J R Statist Soc Ser B 1955;17:69-78.
4. Goodman SN, Hopkins J. Toward evidence-based medical statistics. 1: the P value fallacy. Ann Intern Med 1999;130:995-1004.
5. Ioannidis JPA. Why most published research findings are false. PLoS Med 2005;2:e124. Epub 2005 Aug 30.
6. Rosendaal, FM, Bouter LM. Dwalingen in de methodologie (slot). XXXIX. De ultieme waarheid. Ned Tijdschr Geneeskd 2002;146:304-9.
7. Rosendaal FR. The p-value: a clinician's disease? Eur J Intern Med 2016;35:20-3.
8. Schouten M, Van de Putte E. De (on)zin van screening op kindermishandeling. Huisarts Wet 2018;61:30-4.
9. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. Nat Methods 2015;12:179-85. DOI: 10.1038/nmeth.3288.
10. Szucs D, Ioannidis JPA. When null hypothesis significance testing is unsuitable for research: a reassessment. Front Hum Neurosci 2017;11:390.
11. Trafimow D. Using the coefficient of confidence to make the philosophical switch from a posteriori to a priori inferential statistics. Educ Psychol Meas 2017;77:831-54. Doi:10.1177/0013164416667977.
12. Chavalarias D, Wallach JD, Li AH, Ioannidis JP. Evolution of reporting P values in the biomedical literature, 1990-2015. JAMA 2016;315:1141-8. DOI:10.1001/jama.2016.1952.
13. Trafimow D, Marks M. Editorial. Basic Appl Soc Psych 2015;37:1-2.
14. Valentine JC, Aloe AM, Lau TS. Life after NHST: how to describe your data without 'p-ing' everywhere. Basic Appl Soc Psych 2015;37:260-73.
15. Bos MJ, Buis S. Thirty-minute office blood pressure monitoring in primary care. Ann Fam Med 2017;15:120-3. Doi:10.1370/afm.2041.
16. Van de Schoot R, Denissen J, Neyer FJ, Kaplan D, Asendorpf JB, Van Aken MAG. A gentle introduction to Bayesian analysis: applications to developmental research. Child Dev 2014; 85:842-60. Doi:10.1111/cdev.12169.

De Boer MR, Van Grootel LE, Bouter LM. Stop met het onkritische gebruik van nulhypothesen. Huisarts Wet 2018;61:DOI:10.1007/s12445-018-0255-4.

Vrije Universiteit Amsterdam, Afdeling Gezondheidswetenschappen, Bèta Faculteit, Amsterdam: dr. M.R. de Boer, universitair docent methodologie en toegepaste biostatistiek. Universiteit van Tilburg, Afdeling Methoden & Technieken, Faculteit Sociale & Gedragswetenschappen, Tilburg: dr. L.E. van Grootel, docent/onderzoeker, l.e.vangrootel@uvt.nl. VUmc, Afdeling Epidemiologie en Biostatistiek, Amsterdam: prof. dr. L.M. Bouter, hoogleraar Methodologie en Integriteit.

Mogelijke belangenverstrengeling: niets aangegeven.