Het onderscheidend vermogen van diagnostische tests

J.A. KNOTTNERUS
A. VOLOVICS

Bij een patiënt met hardnekkige hoestklachten vermoedt de huisarts een chronisch obstructieve bronchitis. Ter beoordeling hiervan ausculteert hij de longen. Hij stelt een verlengd expirium vast. Nu vraagt hij zich af, of op grond van deze bevinding een chronisch obstructieve afwijking waarschijnlijker is geworden en of hij hierop het verdere beleid zou moeten afstemmen.

Knottnerus JA, Volovics A. Het onderscheidend vermogen van diagnostische tests [Huisarts, wetenschap en statistiek]. Huisarts Wet 1989; 32(9): 338-46.

Rijksuniversiteit Limburg, Postbus 616, 6200 MD Maastricht.

Prof. dr. J.A. Knottnerus, hoogleraar huisartsgeneeskunde, vakgroep Huisartsgeneeskunde; A. Volovics, vakgroep Medische Informatica en Statistiek.

Correspondentie: Prof. dr. J.A. Knottnerus.

Inleiding

De diagnostiek en de zeeffunctie van de huisarts vormen hoofdbestanddelen van het huisartsenvak. Statistische aspecten van diagnostiek krijgen dan ook veel aandacht in *Huisarts en Wetenschap*. Hierbij gaat het steeds om het verband tussen enerzijds klachten en bevindingen of testuitslagen (samengevat: diagnostica) en anderzijds één of meer mogelijke diagnosen. Er zijn verschillende manieren om dit verband uit te drukken, en daarmee het 'onderscheidend vermogen' van de diagnostiek weer te geven. Indien er geen verband is, is het onderscheidend vermogen nihil, en kan het diagnosticum maar beter niet gebruikt worden. Alleen diagnostica met een goed onderscheidend vermogen zijn geschikt.

Alvorens daar verder op in te gaan, is het wenselijk nog iets te zeggen over hoe hier het begrip 'kans' wordt gebruikt. Hierbij is sprake van toepassing van informatie over een grotere groep personen op individuele gevallen. De 'kans op een ziekte' is dan equivalent aan de prevalentie (relatieve frequentie) van de ziekte in een (aselecte steekproef uit een) omschreven populatie, op relevante kenmerken overeenkomend met het betrokken individu. Bijvoorbeeld: een populatie spreekuurbezoekers met een bepaalde klacht, in aanmerking komend voor diagnostiek. Elk diagnostisch gegeven dat bekend wordt, geeft dan een nadere specificatie van de populatie.

Onderscheidend vermogen

Muris e.a. onderzochten het verband tussen bevindingen bij auscultatie en de Tiffeneau-waarden.[1] Daartoe werd bij 86 patiënten die zich tot de huisarts wendden met klachten die mogelijk verband hielden met een aandoening van de onderste luchtwegen, zowel auscultatie als spirometrie verricht (*tabel 1*). Het onderscheidend vermogen van de bevinding 'verlengd expirium' voor de diagnostiek van bronchusobstructie (met als standaard-test de afwijkende Tiffeneau-waarde) is nu uit te drukken in:

- sensitiviteit: de kans op een afwijkende bevinding bij zieken = 27 procent;
- specificiteit: de kans op niet-afwijkende bevindingen bij niet-zieken = 67 procent.

Van de zieken is dus maar liefst 73 procent fout-negatief, en van de niet-zieken is 33 procent fout-positief.

Het is duidelijk dat een test beter is, naarmate zowel de sensitiviteit als de specificiteit dichter naderen tot 100 procent. Informatie over elk apart heeft op zichzelf niet veel betekenis. Een absurd voorbeeld kan dit verduidelijken:

Vrijwel iedere diabetespatiënt heeft een neus. De sensitiviteit van de test 'het hebben van een neus' voor de diagnostiek van diabetes is dus vrijwel 100 procent. Dat lijkt indrukwekkend, maar personen zónder diabetes hebben niet minder vaak een neus. De specificiteit van dit kenmerk (het percentage niet-zieken zonder neus) is dan uiteraard 0 procent. Men kan het ook zo zeggen: de kans op het hebben van een neus is bij diabetes- en niet diabetespatiënten even groot, dus deze 'test' discrimineert niet.

Dat geldt volgens *Muris e.a.* overigens ook voor het verlengd expirium bij bronchusobstructie: 27 procent van de patiënten met bronchusobstructie heeft een verlengd expririum, vergeleken met 33 procent van de niet-patiënten.[1]

Het is handig om de informatie van sensitiviteit en specificiteit van een testuitslag samen te vatten in één maat voor het onderscheidend vermogen c.q. het verband tussen testuitslag en de aanwezigheid van ziekte. Dat kan via het zogenaamde aannemelijkheidsquotiënt A. Dit is het quotiënt van:

- de kans op (de aannemelijkheid van) een bepaalde testuitslag bij zieken (teller);

en:

- de kans op (de aannemelijkheid van) die uitslag bij niet-zieken (noemer).

In geval van een positieve uitslag (A+) is dit quotiënt equivalent aan: sensitiviteit/(100% − specificiteit); bij een negatieve uitslag (A−) is het: (100% − sensitiviteit)/specificiteit.

Als teller en noemer aan elkaar gelijk zijn, heeft de test in het geheel geen onderscheidend vermogen: A is dan gelijk aan 1,0 en men kan dan net zo goed een munt

opgooien. Deze situatie doet zich voor bij de test 'het hebben van een neus' bij de diagnostiek van diabetes, en wordt in het onderzoek van *Muris e.a.* benaderd door het verlengd expirium bij bronchusobstructie; de bijbehorende aannemelijkheidsquotiënten zijn respectievelijk:

A+ = 100%/(100% − 0) = 1,0;

A− = 27%/(100% − 67%) = 0,82.

Het discriminerend vermogen is beter naarmate A+ groter is dan 1, en naarmate A− meer nadert tot 0,0. Ter oriëntatie hebben wij voor een aantal positieve testuitslagen (en symptomen) de aannemelijkheidsquotiënten berekend (*tabel 2*). We zien een sterk wisselend beeld. De gamma GT scoort nauwelijks hoger dan de test 'munt opgooien' bij het diagnostiseren van levermetastasen. Typische angina pectoris discrimineert zeer goed bij het vaststellen van coronaire pathologie, veel sterker dan atypische angina pectoris, en nog veel sterker dan een moderne test als radionuclide angiografie. Een tussenpositie wordt ingenomen door het sediment bij de diagnostiek van urineweginfecties (met de kweek als standaard) en de diafanoscopie bij de herkenning van sinusitis maxillaris (met de echoscopie als standaard). Hetzelfde geldt voor het vermogen van de huisarts om op grond van het klinisch beeld te voorspellen of er sprake is van een afwijkend ECG.

Bij een test met meer dan twee mogelijke klassen van uitslagen, is het mogelijk om voor elke klasse een aannemelijkheidsquotiënt te berekenen. Dat geldt bijvoorbeeld voor veel laboratoriumuitslagen, zoals het CPK of de bloedsuiker. Bij een dichotome test zijn er maar twee klassen en dus ook maar twee aannemelijkheidsquotiënten: A+ en A−. Het is dan aantrekkelijk om ook A+ en A− weer in één allesomvattende maat voor het onderscheidend vermogen voor positieve én negatieve uitslagen samen te vatten. Het meest in aanmerking hiervoor komt de Odds-ratio: het kruisprodukt $(a \cdot d)/(b \cdot c)$.[7] Uit de vierveldentabel is gemakkelijk af te leiden, dat dit quotiënt gelijk is aan A+/A−.

Bij de weergave van etiologische verbanden wordt de Odds-ratio veel gebruikt; als het om diagnostiek gaat, gebeurt dat veel minder. Toch is deze maat erg handig. Aldus kan men bijvoorbeeld in één oogopslag zien welke van de vier door *Van Duijn* beschreven methoden van diafanoscopie het beste overall-onderscheidend vermogen heeft ten aanzien van de echoscopische standaard (*tabel 3*). Het blijkt nu dat de derde testprocedure niet onderdoet voor de eerste twee, terwijl de laatste veruit het sterkst discrimineert.

Tabel 1 *Verband tussen bevindingen bij auscultatie en Tiffeneau-waarde bij 86 patiënten.*

Auscultatie	Tiffeneau-waarde		
	afwijkend	normaal	totaal
Verlengd expirium	16 (a)	9 (b)	25
Geen verlengd expirium	43 (c)	18 (d)	61
Totaal	59	27	86

Bron *Muris e.a.*[1]

A priori kans:	(a+c)/(a+b+c+d)	= 59/86 = 69%
Sensitiviteit:	a/(a+c)	= 16/59 = 27%
Specificiteit:	d/(b+d)	= 18/27 = 67%
Predictieve waarde		
– verlengd expirium:	a/(a+b)	= 16/25 = 64%
– geen verlengd expirium:	d/(c+d)	= 18/61 = 30%

Tabel 2 *Aannemelijkheidsquotiënten voor diverse positieve testuitslagen en symptomen.*

Test	Ziekte	Uitslag	Aannemelijkheids-quotiënt A
Muntopgooien	alle	kop/munt	1,0/1,0
Auscultatie[1]	bronchus-obstructie	verlengd expirium	0,82
Gamma GT[4]	levermetastase	verhoogd	1,4
Sediment[5]	urineweginfectie	positief	1,7
Voorspelling huisarts[7]	ECG-afwijking	positief	4,0
Diafanoscopie[6]	sinusitis maxillaris		
– infra-orbitaal			2,6
– pupil			1,5
– subjectief			2,3
– subjectief met neusdruppelen			11,1
Anamnese[2]	coronaire stenose	typische agina pectoris	115
		atypische angina pectoris	14
Radionuclide-angiografie[2]	coronaire stenose	positief	3,6

Van prior kans naar posterior kans

Diagnostische informatie wordt verzameld om wijzer te worden. Dat wil zeggen: om te komen van een eerste indruk, via een testuitslag, naar een 'herziene' indruk. In termen van kansen:

- Men gaat uit van een bepaalde beginkans op ziekte(n), die men kent of schat op grond van de kenmerken van de patiënt, vóórdat de test is uitgevoerd. Deze a priori kans drukt uit in welke

mate er sprake is van een diagnostische hypothese.
- Door toevoeging van de informatie van de testuitslag gaat deze a priori kans over in de a posteriori kans (ook wel de predictieve waarde van de testuitslag genoemd).

Voor het onderzoek *Muris e.a.* betekent dit:
- De a priori kans bij de patiënten in het onderzoek is 59/86=69%.
- Het onderscheidend vermogen van de test wordt gegeven door sensitiviteit, specificiteit, en daaruit volgend de aannemelijkheidsquotiënten A+ en A−.

De predictieve waarde van een positieve uitslag is dan:
a/(a+b)=16/25=64%.
En de predictieve waarde van een negatieve uitslag is:
d/(c+d)=18/61=30%.
We zien dat de a priori kans en de predictieve waarde van de positieve uitslag in dit geval nauwelijks van elkaar verschillen. Dat lag ook voor de hand. We wisten immers al dat de auscultatie nauwelijks (en dan nog in de verkeerde richting) discrimineerde, zodat het niet uitmaakt of men het onderzoek doet en wat de uitslag is.

Anders ligt dit bij de diafanoscopie volgens de best discriminerende methode van *Van Duijn*:
- De a priori kans op sinusitis (volgens de echo) in zijn patiëntengroep was: 36/140=26%.
- De predictieve waarde van de positieve uitslag was: 27/34=79%.
- De predictieve waarde van de negatieve uitslag was: 97/106=92%.[5]

Het maakt hier dus wel degelijk veel uit, of men de testuitslag wel of niet tot zijn beschikking heeft: de a priori kans wordt in sterke mate 'herzien' op grond van de testuitslag.

A priori kans en predictieve waarde

Al vaak is benadrukt dat de predictieve waarde van een testuitslag in sterke mate wordt bepaald door de mate van verdenking vooraf c.q. de a priori kans op ziekte.[8-10] Het voorbeeld in *tabel 4* maakt dit nog eens duidelijk. Uitgaande van een sensitiviteit van het mammogram van 90 procent ten aanzien van carcinoom in de mamma, en een specificiteit van eveneens 90 procent, kunnen we voor elke a priori kans telkens via vierveldentabellen berekenen wat de predictieve waarde is van de diverse uitslagen. In *tabel 5* is een voorbeeld gegeven van deze stapsgewijze berekening. Gedemonstreerd wordt hierdoor dat, naarmate de a priori kans op carcinoom hoger wordt,
- de predictieve waarde van een positieve uitslag hoger wordt;
- de predictieve waarde van een negatieve uitslag lager wordt.

Deze conclusie is intuïtief ook begrijpelijk: men zal in verdachte gevallen minder gauw dan in niet verdachte gevallen genoegen nemen met een negatieve uitslag, en dus sterker rekening houden met een fout-negatief resultaat. Anderzijds zal men in niet-verdachte gevallen, zoals bij bevolkingsonderzoek, eerder geneigd zijn rekening te houden met een fout-positief resultaat.

Deze inzichten zijn van groot belang voor het begrijpen van de verschillen in interpretatie van testuitslagen tussen de huisartsgeneeskunde en de specialistische geneeskunde. In het tweede geval is immers de a priori kans op ernstige ziekte vaak hoger, doordat de huisarts al een

Tabel 3 *Samenvatting van het onderscheidend vermogen van tests (in dit geval diafanoscopie) in de Odds-ratio (OR). Per test is eerst de volledige vierveldentabel gegeven, ontleend aan Van Duijn.*[16]

Transilluminatie-procedure	Uitslag	Echo (standaard) +	Echo (standaard) −	A+	A−	OR
Infra-orbitaal	+	21	27	2,6	0,44	6,0
	−	10	77			
Pupil	+	28	64			
	−	3	40	1,5	0,25	5,8
Subjectief	+	27	34	2,3	0,37	6,2
	−	9	70			
Subjectief met neusdruppelen	+	27	7	11,1	0,27	41,6
	−	9	97			

A+ = sensitiviteit/(100% − specificiteit).
A− = (100% − sensitiviteit)/specificiteit.
OR = A+/A− = a·d/b·c.

Tabel 4 *Uitvoering van mammografie in verschillende situaties, uitgaande van constante waarden van sensitiviteit (90%) en specificiteit (90%) (rekenvoorbeeld). Percentages.*

Situatie vrouw	A priori kans	Predictieve waarde + uitslag	Predictieve waarde − uitslag
Jong, klachtenvrij	0,2	1,8	≈100
Middelbaar, klachtenvrij	0,5	4,3	≈100
20-30 jaar, knobbeltje in de borst bij huisarts	5	32,0	99,4
50-60 jaar, knobbeltje in de borst bij huisarts	40	86,7	93,1
70-80 jaar, knobbeltje in de borst bij huisarts	90	98,8	50

selectie heeft gemaakt.[9] Daarnaast wordt duidelijk, hoe men in statistische zin over indicatiegebieden voor tests kan spreken. Zo kan men zich voorstellen dat de predictieve waarde van een positieve uitslag *beneden* een bepaald minimum van de a priori kans zó laag blijft, dat er geen consequenties zijn voor nader onderzoek en therapie. *Boven* een bepaald maximum (bijvoorbeeld bij een oudere vrouw met een mammatumor) is het onderzoek ook niet zinvol, omdat nu de predictieve waarde van een negatieve uitslag zo laag is, dat zelfs bij een negatieve uitslag nader onderzoek (een biopsie) gedaan zal worden. Het indicatiegebied wordt dus in principe begrensd door een minimum en maximum a priori kans (*figuur 1*). Deze 'drempelwaarden' zijn ook afhankelijk van het belang en de ernst van de diverse consequenties.[11]

Tabel 5 *Voorbeeld van de berekening van de predictieve waarde op basis van gegeven waarden over a priori kans, sensitiviteit en specificiteit. Voor het gemak wordt uitgegaan van een populatie van 1000 personen.*

Stap 1 priori kans = 40%

		kanker +	kanker −	
mammografie	+			
	−			
		400		1000

Stap 2 sensitiviteit = 90%

		kanker +	kanker −	
mammografie	+	360		
	−			
		400		1000

Stap 3 specificiteit = 90%

		kanker +	kanker −	
mammografie	+	360		
	−		540	
		400	600	1000

Stap 4 verder invullen tabel en berekening predictieve waarden

		kanker +	kanker −	
mammografie	+	360	60	420
	−	40	540	580
		400	600	1000

Predictieve waarde:
- positieve uitslag: 360/420 = 85,7%
- negatieve uitslag: 540/580 = 93,1%

Bij het bovenstaande is ervan uitgegaan dat het onderscheidend vermogen van een test gelijk blijft als de a priori kans toeneemt. Vaak gaat dit bij benadering op, maar men mag hier niet altijd op rekenen. Zo zal bij de chirurg niet alleen de a priori kans op mammacarcinoom hoger zijn dan in de huisartspraktijk, maar de gepresenteerde tumoren zullen gemiddeld ook groter zijn (verder gevorderd zijn). In zulke gevallen is te verwachten dat de sensitiviteit hoger uitvalt (grotere carcinomen zijn beter te ontdekken) en de specificiteit lager (bij grotere benigne tumoren kunnen meer fout-positieve uitslagen voorkomen), terwijl ook het aannemelijkheidsquotiënt van een positieve uitslag lager kan worden. Daarnaast is er selectie met betrekking tot testuitslagen als gevolg van het verwijsproces.[12] In geval van positieve uitslagen wordt in de regel vaker en eerder verwezen, en alleen dat al maakt na verwijzing de kans op positieve uitslagen groter bij zowel zieken als (uiteindelijk) niet-zieken. In de huisartspraktijk zal daarom het onderscheidend vermogen van veel tests anders zijn dan in de specialistische praktijk.

Nadere uitwerkingen

Het bovenstaande is niet moeilijk te begrijpen, en wie het heeft begrepen kan in principe de diverse nadere uitwerkingen op dit gebied volgen en leren toepassen. Over dergelijke uitwerkingen is veel gepubliceerd, ook in *Huisarts en Wetenschap*.[8-10] We zullen hier nog enkele aspecten kort bespreken.

- Er wordt veel geschreven over de zogenaamde regel van Bayes, als hulpmiddel om uit de a priori kans de a posteriori kans te berekenen. Dit kan handig zijn als men snel te werk wil gaan met behulp van een zakcalculator of als er via een computer in hoog tempo vele berekeningen moeten worden uitgevoerd. In zijn meest basale

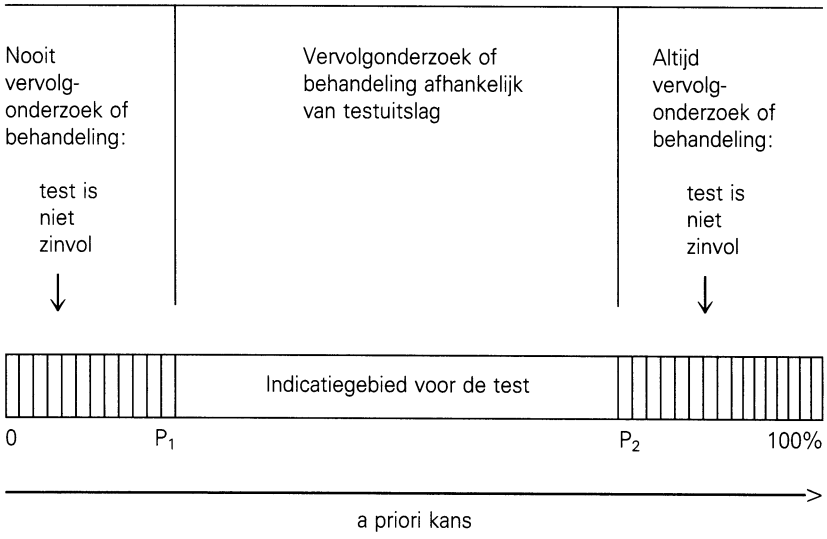
gedaante is de regel van Bayes echter niets anders dan de weergave van de predictieve waarden vanuit een vierveldentabel, in de vorm van een algemene formule. Op deze wijze kan men de regel dan ook gemakkelijk onthouden en zelf reproduceren (*tabel 6*).

- Men kan zich veel rekenwerk besparen door gebruik te maken van een *nomogram*; daarin kan men, uitgaande van de a priori kans en het aannemelijkheidsquotiënt, de a posteriori kans aflezen. Dit nomogram is af te leiden uit de regel van Bayes (*bijlage*).[2 13] Een andere grafische benadering, waarbij men op grond van sensitiviteit, specificiteit en a priori kans gemakkelijk de predictieve waarden kan aflezen, is het 'vierkantsdiagram', ontwikkeld door de Nederlandse huisarts *Baggen*.[14]

- Vaak moet men bij een test waarbij er meer dan twee klassen van uitslagen zijn (bij tests met een ordinale of numerieke testvariabele), beslissen welke uitslagen men als positief c.q. negatief zal beschouwen. Men kan daarvoor een 'afkappunt' kiezen: een drempel tussen positieve en negatieve uitslagen. Bij elk afkappunt hoort dan een eigen combinatie van de waarden van sensitiviteit en specificiteit. In het algemeen geldt: hoe hoger de drempel waarbij men van een positieve uitslag spreekt, des te lager de sensitiviteit (meer fout-negatieve uitslagen) en des te hoger de specificiteit (minder fout-positieve uitslagen). In dergelijke gevallen kan het verband tussen sensitiviteit en specificiteit voor de diverse afkappunten worden weergegeven in een grafiek, de zogenaamde ROC-curve (*figuur 2*). Het voordeel van zo'n curve is, dat het onderscheidend vermogen van een bepaalde test over het gehele traject van mogelijke afkappunten kan worden vergeleken met dat van andere tests.[15]

Er zijn diverse benaderingen beschreven om het 'optimale' afkappunt te definiëren.[15] Deze zijn echter verre van eenvoudig en in de praktijk zelden toepasbaar. Dit komt mede doordat moet worden afgewogen hoe de consequenties van fout-positieve en fout-negatieve uitslagen ten opzichte van elkaar gewaardeerd worden, en dat kan per situatie verschillen.[9] In het

Figuur 1 *Schematische weergave van het indicatiegebied van een test.*



P_1 = de a priori kans waarbij de predictieve waarde van een positieve uitslag nog net hoog genoeg is om vervolgonderzoek of -behandeling te rechtvaardigen.

P_2 = de a priori kans waarbij de predictieve waarde van een negatieve uitslag net laag genoeg is om van vervolgonderzoek of -behandeling af te zien.

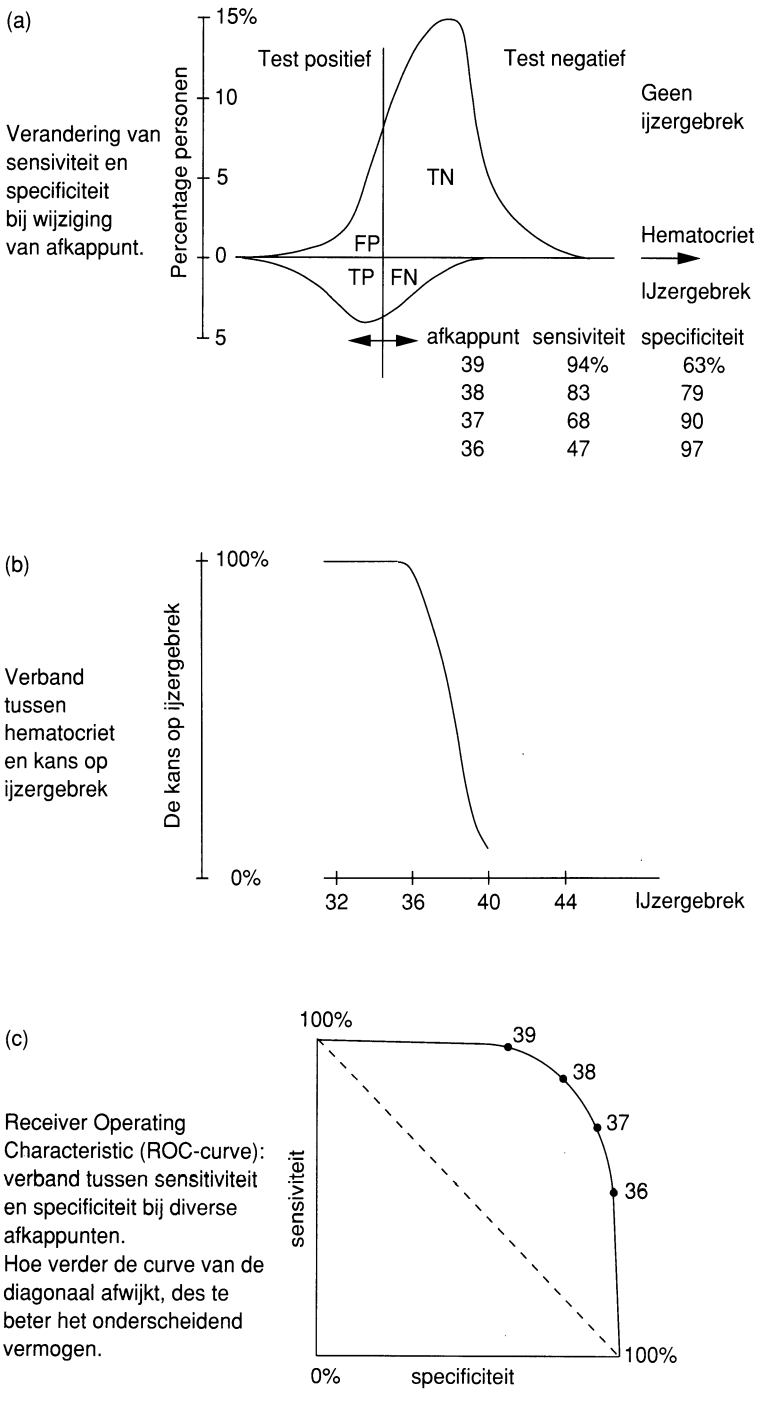
Tabel 6 *Afleiding van de regel van Bayes uit het vierveldendiagram.*

	Ziek	Niet ziek	Totaal
Test +	sensitiviteit × a priori kans (a)	(100% – specificiteit) × (100% – a priori kans) (b)	a + b
Test –	(100% – sensitiviteit × a priori kans (c)	(specificiteit) × (100% – a priori kans) (d)	c + d
Totaal	a priori kans	100% – a priori kans	100%

De regel van Bayes is nu:
- predictieve waarde van positieve uitslag = a/(a+b)
- predictieve waarde van negatieve uitslag = d/(c+d)

waarbij
a = sensitiviteit × a priori kans
b = (100% – specificiteit) · (100% – a priori kans)
c = (specificiteit × (100% – a priori kans)
d = (100% – sensitiviteit) × a priori kans)

Figuur 2 *Relatie tussen hematocriet en ijzergebrek, en de ROC-curve*



algemeen geldt: naarmate het erger is om een ziektegeval te missen, zal de sensitiviteit hoger moeten zijn (minder fout-negatieven). En naarmate de nadelige effecten van behandeling groter zijn, zal de specificiteit hoger moeten zijn (minder fout-positieven).

- Het kiezen van een afkappunt (en het aldus indelen van de testuitslagen in positieve en negatieve uitslagen) bij ordinale of numerieke testvariabelen is nodig indien er een beslissing moet worden genomen (bijvoorbeeld met betrekking tot al dan niet behandelen of verwijzen). Strikt genomen gaat dit indelen van veel verschillende waarden in slechts twee groepen gepaard met informatieverlies. Een sterk verhoogde bloedsuiker zegt immers meer over de aanwezigheid van diabetes mellitus dan een nauwelijks verhoogde waarde, hoewel beide in de klasse 'verhoogd' vallen. Om aan dit informatieverlies te ontkomen, kan men voor alle uitslagen telkens precieze predictieve waarden afleiden, indien men voor die (klasse van) uitslagen voldoende informatie heeft over het al dan niet aanwezig zijn van de ziekte (*figuur 2c*). Het verband tussen testuitslagen en predictieve waarde kan bij statistische analyse van onderzoeksresultaten ook worden uitgedrukt in een regressievergelijking.[16]

- In dit artikel hebben we ons beperkt tot het toepassen van enkelvoudige tests bij enkelvoudige diagnostische hypothesen. Men kan echter dezelfde principes toepassen op situaties waarbij meer diagnostische hypothesen worden getoetst, of waarbij diverse tests worden toegepast. Als er meer dan één test wordt toegepast, moet onder meer rekening worden gehouden met het feit dat in dat geval alleen al op grond van het toeval een grotere kans bestaat op foutpositieve uitslagen.
- Voorts moet men bedacht zijn op zogenaamde afhankelijkheid van testuitslagen: als de ene test eenmaal positief of negatief is uitgevallen, kan het onderscheidend vermogen van andere tests in deze beide categorieën verschillend uitvallen. Dit is begrijpelijk: als eenmaal palpatie van de mamma positief is uitgevallen, is er ook meer kans dat een eventueel aanwezig carcinoom bij mammografie te zien is.

Methodologische aspecten

Het bepalen van het onderscheidend vermogen van tests geschiedt via wetenschappelijk onderzoek. De kern daarvan is dat de te evalueren test en een daarvan onafhankelijk 'gouden' standaard-diagnosticum worden toegepast op alle te onderzoeken personen. Vervolgens wordt de samenhang tussen de resultaten van beide tests bepaald en uitgedrukt in de in het voorgaande besproken parameters voor het onderscheidend vermogen.

Bij dergelijk onderzoek gelden vergelijkbare eisen, en liggen in principe dezelfde bronnen van vertekening op de loer als bij etiologisch of effectiviteitsonderzoek (selectiebias en verstorende variabelen). Wij gaan daarop in deze aflevering niet apart in. Wel zij gezegd dat bij veel, ook reeds in zwang zijnde tests, dergelijk onderzoek onvoldoende is gedaan en dat over het onderscheidend vermogen derhalve te weinig bekend is. Er is dus veel werk aan de winkel op dit gebied. Dat geldt speciaal voor de huisartspraktijk, waarop resultaten van tweedelijns onderzoek niet zonder meer van toepassing zijn: niet alleen de a priori kans, maar ook de mate van ontwikkeling van ziekten is immers geheel verschillend. In de huisartspraktijk lijken de zieken op gezonden, en bij de specialist lijken de gezonden op zieken (anders waren zij niet verwezen).[9] Dit stelt in beide gevallen speciale eisen aan de te gebruiken diagnostica. Een bijzondere eis in de huisartspraktijk is bovendien, dat er tests zijn met een breed toepassingsgebied (zoals de BSE), dat wil zeggen: een onderscheidend vermogen voor meer dan één ziekte. De 'probleemruimte' is immers nog groot.

Een centraal vraagstuk is dat van de 'gouden standaard'. Het is zeker bij eerstelijns onderzoek lang niet altijd mogelijk zo'n standaard-diagnosticum toe te passen. Het kan zijn dat zo'n standaard te invasief is of te complex om te gebruiken, bijvoorbeeld als het gaat om de evaluatie van het onderscheidend vermogen van chronische buikklachten ten opzichte van een 'standaard-pakket' van foto's, scopieën en leverfunctietests. Een ander probleem is dat er vaak geen onafhankelijk en eenduidig standaard-diagnosticum bestaat, zoals bij 'minimal brain dysfunction', migraine, lage-rugklachten of zelfs bij sinusitis maxillaris. In het laatste geval is onder meer de vraag of een positieve kweek van het kaakholtepunctaat of de echo-uitslag wel equivalent zijn aan de sinusitis. In deze gevallen moeten andere criteria worden gekozen, zoals het klinisch beloop of een uitspraak van een panel van experts.

Bij kritische beschouwing blijft er ook bij zogenaamde 'harde' diagnosen en diagnostica weinig over van het begrip 'gouden' standaard. Zelfs een röntgenfoto, pathologisch anatomisch preparaat of een CT-scan levert geen perfect onderscheidend vermogen op, en laat ruimte voor interwaarnemer-variaties. Het gaat dus in werkelijkheid veeleer om het maken van goede controleerbare afspraken over acceptabele diagnostische criteria, dan om het vinden van de ideale gouden standaard.

Een belangrijk methodologisch vereiste is, dat uitslagen van alle onderzochte personen in de analyse worden betrokken. Indien men bijvoorbeeld personen met twijfelachtige of onduidelijke testuitslagen of met een onduidelijke ziektestatus hieruit weglaat, zal men geflatteerde waarden vinden voor sensitiviteit, specificiteit en predictieve waarden. Men dient voor dergelijke 'onduidelijke' bevindingen aparte categorieën in de tabel te reserveren.

Bij diagnostisch wetenschappelijk onderzoek geldt, evenals bij ander onderzoek, dat er een voldoende grote onderzoeksgroep moet zijn om nauwkeurige uitspraken te kunnen doen. Om die reden moet ook in dit geval de gewenste steekproefgrootte zoveel mogelijk vooraf worden geschat.[17] Tevens is het aan te bevelen om bij de gevonden resultaten van sensitiviteit, specificiteit, aannemelijkheidsquotiënt of odds-ratio een 95%-betrouwbaarheidsinterval te vermelden. Dit wordt meestal niet gedaan, en dan heeft men onvoldoende inzicht in de nauwkeurigheid van de bevindingen. Ter illustratie zijn 95%-betrouw-

Tabel 7 *95%-betrouwbaarheidsintervallen van sensitiviteit, specificiteit en aannemelijkheidsquotiënten van een tweetal tests.*

	Verlengd expirium t.a.v. bronchus-obstructie[1] n = 86	Diafanoscopie (subjectie lichtperceptie met neusdruppelen[5]) n = 140
Sensitiviteit	21% – 33%	69% – 82%
Specificiteit	58% – 76%	91% – 96%
A+	0,48 – 1,4	6,2 – 19,7
A–	0,87 – 1,4	0,20 – 0,37
Odds-ratio	0,28 – 2,0	17,1 – 101,4

Tabel 8 *Methodologische manco's in gepubliceerd onderzoek bij 'ijking' van diagnostische tests. Percentages van het aantal onderzoeken.*

Geen 'gouden' standaard	32
Definitie positief/negatief onduidelijk	32
Geen 'blinde' observaties	60
Gegevens onduidelijk gepresenteerd	52
Fout of geen gebruik van termen:	
– sensitiviteit	21
– voorspellende waarde	69
Geen rekening gehouden met setting of a priori kans	81

Bron *Sheps and Schechter.*[17]

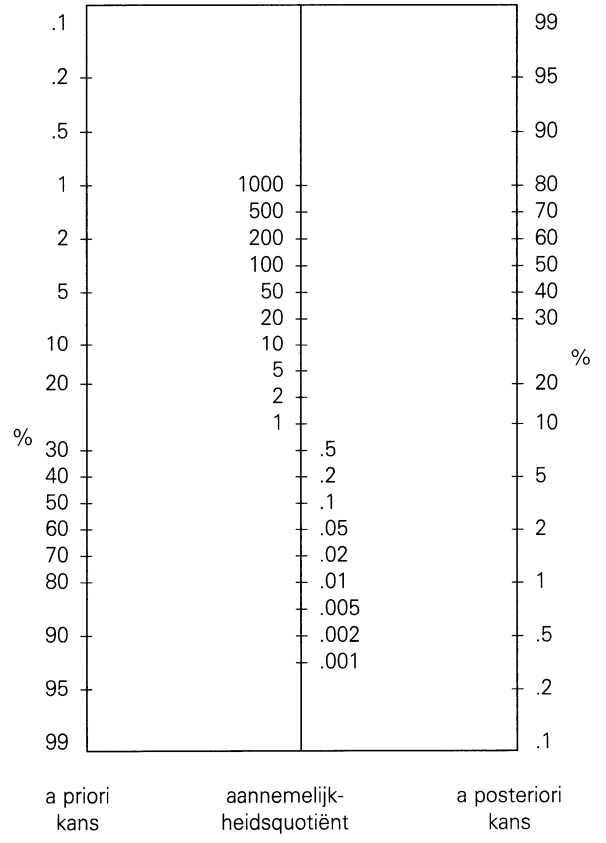
baarheidsintervallen weergegeven voor resultaten van *Muris* en *Van Duijn* (*tabel 7*).

Uiteraard verdient ook de reproduceerbaarheid van de test de aandacht. Een goede reproduceerbaarheid is een voorwaarde voor een goed onderscheidend vermogen, en is een speciaal punt van zorg bij tests die bij herhaling bij dezelfde personen moeten worden uitgevoerd. Voor het in kaart brengen van deze reproduceerbaarheid kan men gebruik maken van methoden zoals in de vorige aflevering zijn besproken.[18]

Het is niet eenvoudig om in alle opzichten bevredigend wetenschappelijk onderzoek te doen naar de waarde van diagnostica. Een rol speelt al, dat onderzoekers (en auteurs) lang niet altijd op de hoogte zijn van essentiële begrippen. Dit blijkt onder meer uit een inventarisatie die *Sheps et al.* hebben gemaakt van een groot aantal publikaties op dit gebied.[19] De meeste studies vertoonden wel één of meer gebreken (*tabel 8*). Zo'n bevinding is een reden te meer om als lezer extra alert te zijn ten aanzien van het gebodene.

Het kan verhelderend zijn zich de overeenkomst te realiseren tussen diagnostiek en wetenschappelijk onderzoek. Men kan immers een wetenschappelijk onderzoek zien als een 'diagnostische test' om vast te stellen of een bepaalde hypothese wel of niet verworpen moet worden. De sensitiviteit van het onderzoek is dan equivalent aan de 'power' van het onderzoek om een eventueel aanwezig verband op te sporen, en de 'specificiteit' is de kans dat een geldende nulhypothese ook als zodanig wordt herkend (c.q. niet wordt verworpen).[18]

Bijlage *Nomogram voor de relatie tussen a priori kans, aannemelijkheidsquotiënt A en a posteriori kans.*



Toelichting De a posteriori kans is te vinden door met een lineaal vanaf de betreffende a priori kans (pre-test waarschijnlijkheid) via de relevante waarde van het aannemelijkheidsquotiënt de as van de a posteriori kans (post-test waarschijnlijkheid) te snijden.[2]

Beschouwing

Er blijkt een groot assortiment te zijn aan manieren om het verband tussen de uitslagen van een diagnosticum en de aanwezigheid van ziekte uit te drukken. Elk van deze manieren heeft eigen voordelen en beperkingen, en welke men gebruikt, hangt af van het na te streven doel. Bij het analyseren van het onderscheidend vermogen van dichotome testvariabelen ten aanzien van het voorkomen van één bepaalde ziekte levert de viervelden-tabel directe toegang tot de begrippen sensitiviteit, specificiteit, predictieve waarden en theorema van Bayes. De aannemelijkheidsquotiënten en de odds-ratio zijn zeer geschikt om in deze situatie het onderscheidend vermogen van de test samen te vatten. Daarnaast biedt het werken met het aannemelijkheidsquotiënt en een nomogram de mogelijkheid veel rekenwerk over te slaan bij het bepalen van de predictieve waarden.

Bij tests met ordinale of numerieke testvariabelen is het van belang dat men zich rekenschap geeft van de inverse samenhang van sensitiviteit en specificiteit bij diverse afkappunten. De ROC-curve vergemakkelijkt daarbij een vlotte beoordeling van het onderscheidend vermogen, ook in vergelijking met andere tests. In de klinische realiteit zal men niet altijd met afkappunten werken, maar voor alle verschillende (klassen van) uitslagen de predictieve waarden zo precies mogelijk willen weten. Ook dit is in principe mogelijk, indien men kan beschikken over voldoende onderzoeksgegevens. Het is aan te bevelen de oorspronkelijke gegevens steeds weer te geven, zodat de lezer eventueel zelf de door hem gewenste maten of afkappunten kan berekenen.

Het onderzoek bevat diverse methodologische valkuilen en bij het lezen van publikaties is daarom een kritische blik gewenst.

Tenslotte: in het bovenstaande zijn technieken vermeld die in principe exacte berekeningen met zich meebrengen. Of deze exacte berekeningen mogelijk zijn, hangt af van (de kwaliteit van) de beschikbare onderzoeksgegevens. Of ze ook steeds gewenst of bruikbaar zijn, hangt af van de werkwijze van de arts. Vaak werkt en denkt deze niet in exacte percentages en getallen, maar meer in globale categorieën. Enerzijds kunnen de geschetste benaderingen hierop wellicht meer gaan inspelen, anderzijds bieden zij concepten, principes en technieken, die bij de hedendaagse kwaliteitsbeoordeling van diagnostiek niet meer zijn weg te denken.

Literatuur

1 Muris JWM, Vaessen MHJ, Sturmans F. De waarde van auscultatie bij de diagnostiek van bronchusobstructie in de huisartspraktijk. Huisarts Wet 1987; 30: 272-4.

2 Griner PF, Panzer RJ, Greenland P. Clinical diagnosis and the laboratory. Logical strategics for common medical problems. Chicago, Ill.: Year Book Medical Publishers, 1986.

3 Nijpels G, Walig C. Urineweginfecties in een huisartspraktijk. De betrouwbaarheid van diagnostiek en de effectiviteit van de antimicrobiële behandeling. Huisarts Wet 1988; 31: 337-8.

4 Ekering JH. Het elektrocardiogram: een zinvolle uitbreiding van het diagnostisch arsenaal van de huisarts? Huisarts Wet 1986; 29: 236-8.

5 Van Duijn NP. Diafanoscopie van de sinus maxillaris. Huisarts Wet 1987; 30: 268-71.

6 Sackett DL, Haynes RB, Tugwell P. Clinical epidemiology. A basic science for clinical medicine. Boston, Mass.: Little Brown and Company, Boston, 1985.

7 Knottnerus JA, Volovics A. Associatiematen voor etiologische en prognostische verbanden. Huisarts Wet 1987; 30: 391-3.

8 Van der Velden HGM. Diagnose of prognose. Huisarts Wet 1983; 26: 125-8.

9 Knottnerus JA. Interpretatie van diagnostische gegevens. Huisarts Wet 1983; 26: 363-8.

10 Van Geldrop WJ, Van den Bosch JSG. De keuze van diagnostische tests, een besliskundige analyse. Huisarts Wet 1986; 29: 378-81.

11 Knottnerus JA, Leffers P. De invloed van verwijsgedrag op het onderscheidend vermogen van diagnostische tests. Tijdschr Soc Gezondheidsz 1987; volume: pagina's.

12 Knottnerus JA. De betekenis van pijn op de borst voor de diagnostiek van coronaire hartziekte en myocardinfarct. Huisarts Wet 1987; 30(suppl 11): 46-51.

13 Knottnerus, JA. Interpretatie van diagnostische gegeven [Dissertatie]. Maastricht: Rijksuniversiteit Limburg, 1986.

14 Baggen J, Leffers P, Van Leeuwen YD. Het testdiagram: een visueel middel bij de interpretatie van testuitslagen. Ned Tijdschr Geneeskd 1984; 128: 1656-9.

15 Weinstein MC, Fineberg HV. Clinical decision analysis. Philadelphia: Saunders, 1980.

16 Knottnerus JA, Volvovics A. Correlatie en regressie. Huisarts Wet 1988; 31: 18-22.

17 Knottnerus JA, Volvovics A. Statistisch toetsen. Huisarts Wet 1988; 31: 100-5.

18 Knottnerus JA, Volvovics A. Overeenstemming tussen beoordelaar. Huisarts Wet 1989; 32: 56-61, 73.

19 Sheps SB, Schechter MT. The assessment of diagnostic tests. A survey of current medical research. JAMA 1984; 252: 2418-22. ■