De ROC-curve voor de eerste evaluatie van nieuwe diagnostische tests

Y.T. VAN DER SCHOUW
A.L.M. VERBEEK
J.H.J. RUIJS

Bij het evalueren van een diagnostische test aan de hand van sensitiviteit en specificiteit – bedoeld als hulpmiddel bij het stellen van de diagnose – speelt het gekozen afkappunt een overwegende rol. Bovendien kunnen door selectie bij één en hetzelfde afkappunt verschillende waarden voor sensitiviteit en specificiteit worden gevonden. Daarom verdient het aanbeveling een hele range van sensitiviteits- en specificiteitswaarden bij verschillende afkappunten te geven. Deze range kan overzichtelijk worden weergegeven in een Receiver Operating Characteristic (ROC) curve. Het oppervlak onder de ROC-curve, de Area Under the Curve (AUC), is een maat voor het discriminerend vermogen van een test, onafhankelijk van de gekozen afkappunten. Met behulp van de AUC kunnen tests op eenvoudige wijze met elkaar worden vergeleken. Bovendien kunnen de diagnostische prestaties van verschillende beoordelaars van tests of de vorderingen van één beoordelaar met behulp van de AUC met elkaar worden vergeleken.

Van der Schouw YT, Verbeek ALM, Ruijs JHJ. De ROC-curve voor de eerste evaluatie van nieuwe diagnostische tests. Huisarts Wet 1992; 35(5): 204-8.

Vakgroep Medische Informatiekunde en Epidemiologie, Katholieke Universiteit Nijmegen, Verlengde Groenestraat 75, 6525 EJ Nijmegen.

Ir. Y.T. van der Schouw, epidemioloog; Dr. A.L.M. Verbeek, arts-epidemioloog; Prof. dr. J.H.J. Ruijs, radiodiagnost, Instituut voor Radiodiagnostiek, Academisch Ziekenhuis St. Radboud.

Correspondentie: Ir. Y.T. van der Schouw.

Inleiding

Sensitiviteit en specificiteit van diagnostische tests zijn belangrijke grootheden bij het stellen van een diagnose; men kan hiermee voor patiënten met een positieve testuitslag uitrekenen hoe groot de kans is dat zij een bepaalde ziekte hebben.[1-4] Voor het *evalueren* van diagnostische tests is het gebruik van sensitiviteit en specificiteit echter minder geschikt; bij één en dezelfde test worden immers nogal eens heel verschillende waarden voor deze testkarakteristieken gevonden.[5] Dat is onder meer een gevolg van het feit dat voor een positieve testuitslag verschillende afkappunten worden gehanteerd. Ook kunnen de verdelingen van de testuitslagen bij zieken en niet-zieken per studie variëren als gevolg van selectie.

Voor het evalueren van diagnostische tests is het belangrijk over een maat te kunnen beschikken die onafhankelijk is van afkappunten of selectie. De Receiver Operating Characteristic (ROC) curve – een eenvoudig interpreteerbaar plaatje – is zo'n maat.[6 7] Pas als het diagnostisch vermogen voldoende gunstig wordt beoordeeld, dient het geschikte afkappunt te worden bepaald. De constructie en het gebruik van deze ROC-curves wordt in deze bijdrage besproken aan de hand van een onderzoek naar de etiologische relatie tussen de ijzerstatus en de aanwezigheid van een eerste acuut myocardinfarct (AMI).

Het onderzoek

In 1986-1988 werd in Rotterdam een patiënt-controle onderzoek uitgevoerd naar de etiologische relatie tussen de ijzerstatus (serumijzer, ferritine, transferrine en ijzerverzadiging van transferrine) en de aanwezigheid van een eerste acuut myocardinfarct (AMI).[8] In totaal 84 patiënten voldeden aan de insluitcriteria (*bijlage*) en via de burgerlijke stand werden evenveel voor leeftijd (vijfjaars-categorieën) en geslacht gematchte controles geworven. Bij de patiënten werden bloedmonsters genomen bij opname, dat wil zeggen binnen vier uur na het begin van de klachten. Bij de controles werd thuis bloed afgenomen. Serumferritine werd bepaald volgens de enzym-immunoassay methode van *Miles et al*;[9] serumijzer werd bepaald met de methode van *Eskelinen et al.*[10]

In het onderzoek bleek dat bij de AMI-patiënten in de acute fase ferritine vrijkwam en daarna in verhoogde concentraties in het serum aanwezig was, terwijl dat in de controlegroep niet het geval was. Naar aanleiding hiervan werd geopperd dat ferritine van diagnostische waarde zou kunnen zijn bij een mogelijk eerste myocardinfarct.

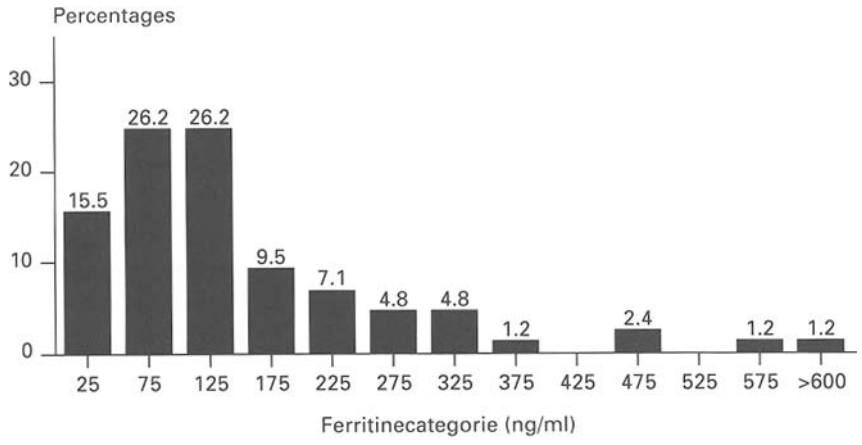
Verdeling van testuitslagen

Patiënten met een bepaalde ziekte zullen niet allen dezelfde uitslag hebben van een bepaalde diagnostische test: sommige uitslagen komen vaak voor, andere – bijvoorbeeld extreem hoge of lage – minder frequent (*figuur 1*).[1] Voor personen zonder de specifieke aandoening (de niet-zieken) kan men een vergelijkbare frequentieverdeling van de testuitslagen maken. Wanneer de twee verdelingen van de testuitslagen weinig overlap vertonen, kunnen we spreken van een goede diagnostische test.

In de praktijk treedt overlap tussen deze twee verdelingen bij nagenoeg alle diagnostische tests op als gevolg van de natuurlijke spreiding van biologische substraten van de onderzochte personen. Ook in het serum van gezonde mensen komt ferritine voor. Wanneer echter blijkt dat het serum van AMI-patiënten een sterk verhoogde hoeveelheid ferritine bevat, is de ferritinebepaling een geschikt diagnosticum.

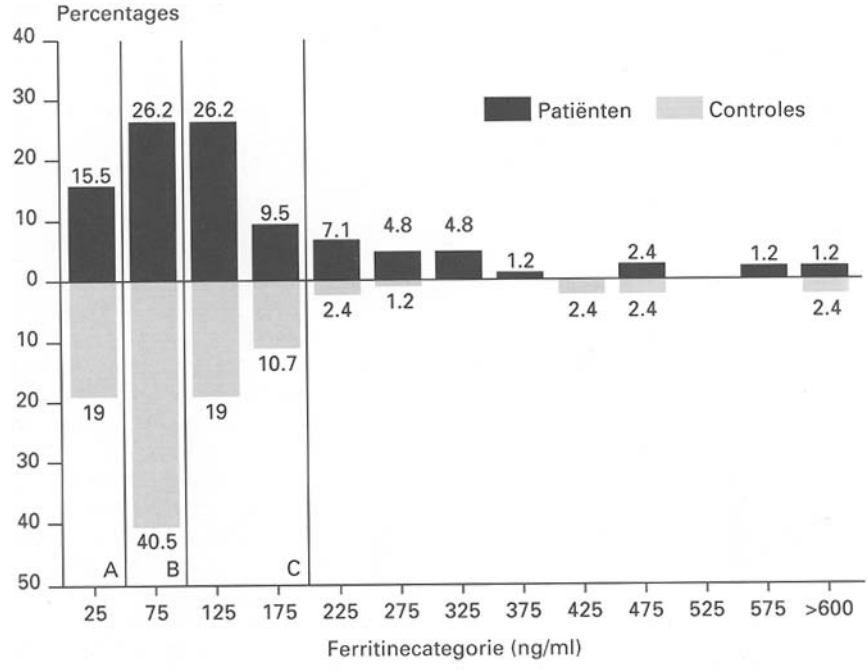
De overlap van de verdelingen van testuitslagen is verder afhankelijk van de samenstelling van de onderzoekspopulatie.[11] Stel dat men scintigrafie met technetium-pyrofosfaat wil evalueren als diagnosticum voor het hartinfarct. In het ene geval bestaat de onderzoeksgroep uit patiënten met de typische anamnese van een myocardinfarct en met ECG-bevindingen die duidelijk passen bij een transmuraal infarct, benevens een duidelijke stijging van de MB-fractie van het creatinefosfokinasegehalte (MB-CKF) in het serum, en de controlegroep uit geneeskundestudenten. In dit geval zal weinig overlap optreden in de scintigrafie-uitslagen. Daarentegen zal een grote overlap optreden, als de onderzoeksgroep

Figuur 1 *Relatieve frequentieverdeling van de 84 patiënten met een eerste acuut myocardinfarct, naar ferritinecategorie.*



De cijfers op de X-as geven het midden van de categorie weer; de eerste categorie loopt dus van 0 t/m 49 enz.

Figuur 2 *Relatieve frequentieverdeling van respectievelijk de 84 patiënten en de 84 controles naar ferritinecategorie.*



A afkappunt 50 ng/ml; sensitiviteit 85%, specificiteit 19%
B afkappunt 100 ng/ml; sensitiviteit 58%, specificiteit 60%
C afkappunt 200 ng/ml; sensitiviteit 23%, specificiteit 89%

bestaat uit patiënten met een acuut myocardinfarct zonder pathologische Q-toppen en slechts een lichte verhoging van het MB-CKF-gehalte, en de controlegroep uit patiënten met oude infarcten en een instabiele angina pectoris.

Afkappunten

In de klinische, diagnostische praktijk hanteert men voor diagnostische tests een afkappunt, bijvoorbeeld een 'normaalwaarde', waaronder een testuitslag negatief wordt genoemd. In de andere gevallen is de test positief. Dit dichotomiseren is bijna inherent aan het handelen van de arts: er wordt verder gediagnostiseerd of niet; er wordt behandeld of niet.[12]

Wanneer voor ferritine een afkappunt van 50 ng/ml wordt gehanteerd, heeft 85 procent van de AMI-patiënten een positieve testuitslag. Dit percentage is de sensitiviteit van de test. Overigens blijkt ook 81 procent van de controles een positieve testuitslag te hebben; men spreekt van 81 procent fout-positieve uitslagen of van een specificiteit van 19 procent. In *figuur 2* is de volledige verdeling van ferritine weergegeven voor beide groepen.

Wanneer een ander afkappunt wordt gekozen, resulteert dit in andere waarden voor de sensitiviteit en de specificiteit van ferritine ten aanzien van het AMI. Als het afkappunt wordt verschoven van 50 ng/ml naar 100 ng/ml of zelfs naar 200 ng/ml, stijgt de specificiteit naar respectievelijk 60 en 89 procent, maar daalt de sensitiviteit naar respectievelijk 58 en 23 procent.

ROC-curve

Een oplossing voor dit probleem van andere waarden voor sensitiviteit en specificiteit bij andere afkappunten, is het geven van de hele *range* van deze testkarakteristieken bij de verschillende afkappunten. Men kan vervolgens uit deze getallen een curve samenstellen, de Receiver Operating Characteristic (ROC) curve.[12-16] Hierbij wordt de sensitiviteit afgezet op de Y-as en het percentage fout-positieve testuitslagen (het complement van de specificiteit) op de X-as.

Voor ferritine bij het AMI gaat de con-

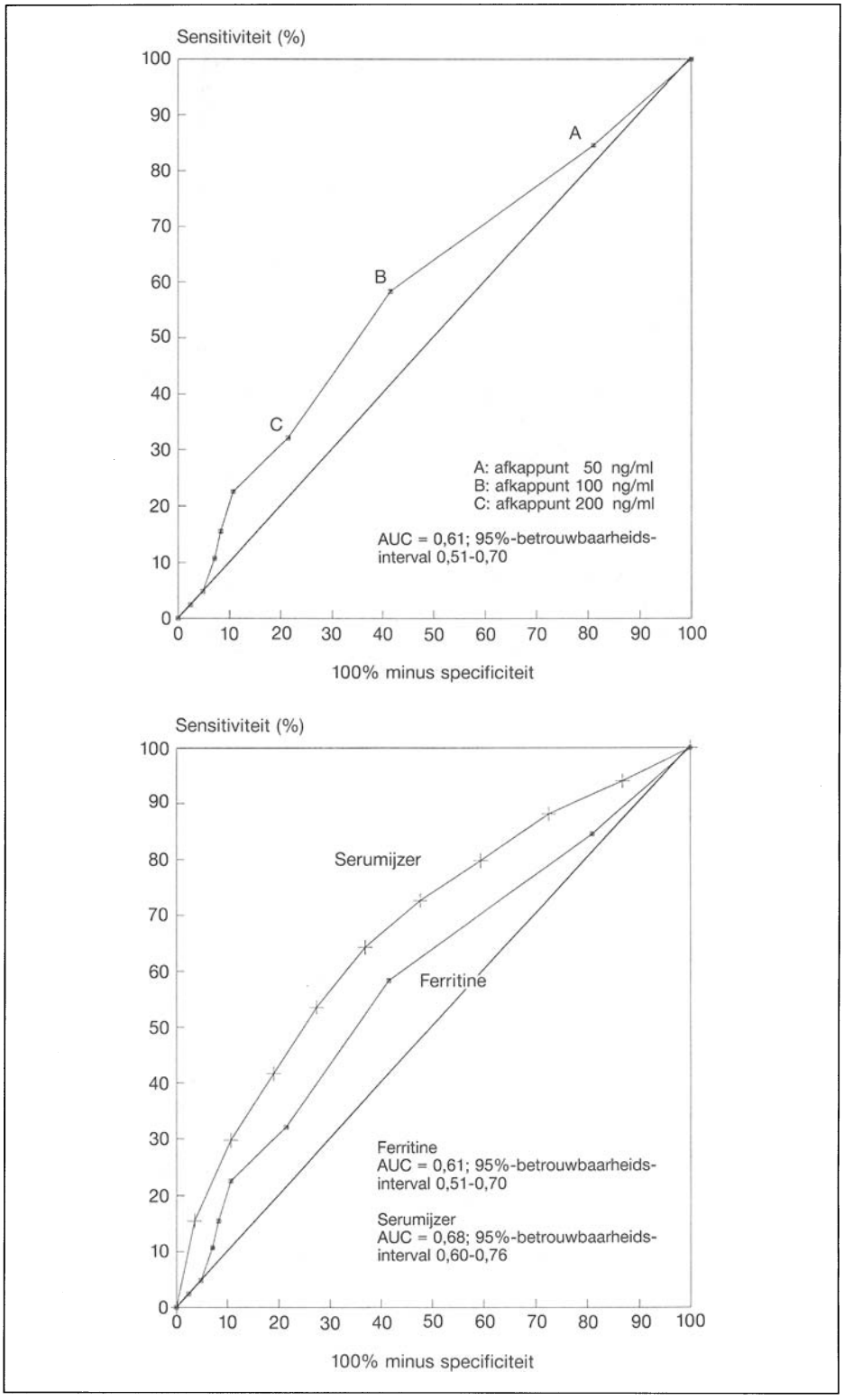
structie van een ROC-curve als volgt. Voor elk van de afkappunten A, B en C in *figuur 2* is de bijbehorende sensitiviteit en specificiteit berekend. Deze uitkomsten zijn uitgezet in *figuur 3*.

Bij een slechte test, waarbij de verdelingen van de uitslagen voor zieken en niet-zieken elkaar volledig overlappen, is op elk afkappunt de sensitiviteit gelijk aan het complement van de specificiteit. De resulterende ROC-curve is dan gelijk aan de diagonaal. Bij een zeer goede test overlappen de verdelingen van testuitslagen zieken en niet-zieken elkaar nauwelijks; dan is de ROC-curve duidelijk verschoven in de richting van de linker bovenhoek van het diagram.[12]

ROC-curven tonen het discriminerend vermogen van een diagnostische test: hoe beter dit is, des te verder is de ROC-curve naar boven en naar links in het diagram verschoven. Het oppervlak onder de ROC-curve (Area Under the Curve = AUC) is een maat voor het discriminerende vermogen van een test. Bij een slechte test is de AUC gelijk aan 0,5. Voor een perfecte test is de AUC gelijk aan 1.*

De AUC van ferritine blijkt 0,61 te zijn met een 95%-betrouwbaarheidsinterval van 0,51-0,70. Dit is bepaald laag en het blijkt dus zinloos om ferritine te bepalen bij patiënten met verdenking op een infarct.

Met behulp van ROC-curves is in één oogopslag te zien of een nieuwe test een beter diagnostisch vermogen heeft dan bestaande tests. In ons patiënt-controle onderzoek is behalve ferritine tevens de serumijzerconcentratie bepaald. Uit *figuur 4* blijkt dat serumijzer een iets beter diagnostisch vermogen heeft dan ferritine, maar toch ook een matig diagnosticum



* De AUC kan worden berekend met behulp van de 'trapezium-methode'.[17] Een diskette met software hiervoor is te verkrijgen via de auteurs.

Figuur 3 (boven) *ROC-curve voor ferritine in serum van 84 patiënten en 84 controles.*

Figuur 4 (onder) *ROC-curves voor ferritine en ijzer in serum van 84 patiënten en 84 controles.*

voor een myocardinfarct is; de AUC is 0,68, met een 95%-betrouwbaarheidsinterval van 0,60-0,76.

Het is mogelijk statistisch te toetsen of de verschillen in de AUC's betekenis hebben. Dit is uitgewerkt door *Hanley*.[18]

Tests met een positieve/ negatieve uitslag

Ook bij tests met alleen een positieve of negatieve uitslag – zoals de screeningsmammografie bij het landelijk bevolkingsonderzoek naar borstkanker[19] – is de constructie van een ROC-curve mogelijk. De radiodiagnost geeft dan met een percentage weer hoe zeker hij is van een uitslag, bijvoorbeeld 100 procent als hij volledig zeker is van de aanwezigheid van een maligne tumor, en 0 procent als met zekerheid gezegd kan worden dat er geen maligniteit aanwezig is. De aldus gescoorde percentages kunnen worden ingedeeld in categorieën, waarna de ROC-curve kan worden geconstrueerd.

Niet iedere beoordelaar zal op dezelfde wijze een positieve of negatieve testuitslag geven. Met behulp van ROC-curves kan worden nagegaan of er beoordelingsverschillen zijn tussen onderzoekers en hoe groot deze verschillen zijn. Discrepanties tussen beoordelaars – inter-observervariatie – kunnen het gevolg zijn van een werkelijke discrepantie in het beoordelingsvermogen. Ook bestaat de mogelijkheid dat de ene onderzoeker foto's consequent met een iets hoger percentage zekerheid beoordeelt dan een andere onderzoeker. Hierdoor stijgt de sensitiviteit van de eerste onderzoeker, maar dat gaat dan ten koste van de specificiteit. De punten van deze persoon liggen op dezelfde ROC-curve als die van de ander, alleen zijn ze naar rechts verschoven.

Verder treedt ook nog intra-observervariatie op. Als een onderzoeker meer ervaring krijgt in het beoordelen van testresultaten, zal zijn ROC-curve gaan verschuiven in de richting van de linker bovenhoek. Bij het Landelijk Referentiecentrum Borstkankerscreening wordt dan ook gebruik gemaakt van ROC-curves om vorderingen van screeningsradiologen na te gaan tijdens bijscholingen.[20]

Beschouwing

Door selectie kunnen de verdelingen van testuitslagen bij zieken en niet-zieken variëren per onderzoek. Dit levert in elk geval verschillende waarden voor sensitiviteit en specificiteit op bij dezelfde afkappunten. De ROC-curve lijkt tamelijk ongevoelig voor deze selectie.[6 7] De vorm van de ROC-curve en de AUC blijven meestal ongeveer gelijk, terwijl de afzonderlijke punten van sensitiviteit en specificiteit verschuiven ten opzichte van de ongeselecteerde situatie.

Pas als uit gedegen evaluatie is gebleken dat een test geschikt is voor gebruik in de klinische praktijk, is het de moeite waard om het geschiktste afkappunt voor de test te bepalen. Dit afkappunt zal variëren, afhankelijk van de situatie waarin de test wordt gebruikt. In een specialistische praktijk zullen andere eisen aan een test worden gesteld dan in een huisartspraktijk.

Bij de evaluatie van een nieuwe diagnostische test kan men zich – net als bij geneesmiddelen-evaluatie – een gefaseerde procedure voorstellen. Daarbij dient steeds aandacht te worden besteed aan het selecteren van een juiste patiëntpopulatie.[21]

Voor een eerste evaluatie kan worden volstaan met een 'makkelijke' populatie, bijvoorbeeld bestaande uit gezonde mensen en mensen met de te diagnostiseren ziekte. De ROC-curve van de nieuwe test wordt bepaald. Als de AUC in deze 'makkelijke' onderzoeksgroep onvoldoende gunstig is, wordt de evaluatie gestopt: blijkbaar is de nieuwe test het ontwikkelingsstadium nog niet te boven.

Men kan zich ook meteen richten op de patiëntenpopulatie waarvoor de test is bedoeld: personen met een verdenking op de aanwezigheid van een bepaalde ziekte. Hierbij dient een aselecte steekproef uit het gehele domein van de in aanmerking komende patiënten – de differentieel-diagnostische patiëntengroep – in het onderzoek te worden betrokken.

In de eerste fase kan men ook doelbewust zoeken naar het best metende diagnostische kenmerk. Ferritine en serumijzer bleken in ons onderzoek weinig voor elkaar onder te doen als diagnosticum voor het AMI, maar wellicht levert het transferrine-gehalte of een samengestelde maat als de ijzerverzadiging van transferrine een hogere diagnostische waarde op.

Wanneer de resultaten van de diagnostische test in de eerste fase bevredigend zijn (hoge AUC), volgt de fase van de plaatsbepaling van de nieuwe test binnen het bestaande diagnostische arsenaal. Naast de uitslag van de nieuwe diagnostische test bij de doelpopulatie is in deze fase ook de informatie van alle andere diagnostische tests van belang: anamnese, fysische diagnostiek, routine-laboratoriumtests, specifiek röntgenonderzoek, functietests, enz. Het diagnostisch vermogen van de onderscheiden tests wordt bepaald met multivariate statistische technieken.[22]

Dan is ook het moment aangebroken voor een 'medical technology assessment'. Wellicht is het mogelijk bepaalde tests – gepaard gaande met invasief onderzoek, hoge kosten of gevaar van complicaties – achterwege te laten zonder dat de diagnostiek tekort wordt gedaan.

Gelet op de huidige stroom van nieuwe diagnostische tests is het redelijk deze drie onderzoeksfasen te doorlopen alvorens een nieuwe test toe te laten, dan wel te laten registreren. De ROC-analyse doet met name dienst in de eerste fase.

Bijlage. De gehanteerde insluitcriteria

Patiënten
- verhoogde serum creatinekinase-spiegel;
- verhoogde serum aspartaat-aminotransferase-spiegel;
- specifieke ECG-afwijkingen;
- jonger dan 75 jaar;
- maximaal 6 maanden voor infarct hartklachten;
- geen chronische hartafwijkingen;
- gebruik van normale voeding;
- niet onder behandeling voor nier of longaandoeningen;
- niet onder behandeling voor alcohol of drugsmisbruik.

Controles
- niet bekend met hartaandoeningen;
- gebruik van normale voeding;
- niet onder behandeling voor nier of longaandoeningen;
- niet onder behandeling voor alcohol of drugsmisbruik.

Dankbetuiging

Met dank aan Dr. Ir. F. J. Kok, Instituut Toxicologie en Voeding-TNO, Zeist en Dr. E. G. Schouten, arts, Vakgroep Humane Epidemiologie en Gezondheidsleer, LU Wageningen, voor het beschikbaar stellen van de gegevens, en aan de studenten van de module Klinische Epidemiologie KUN 1990 voor hun commentaar op eerdere versies.

Literatuur

1 Van der Helm HJ, Hische EAH. Gevoeligheid, specificiteit en diagnostische waarde van laboratoriumonderzoekingen. Ned Tijdschr Geneeskd 1979a; 123: 1944-51.

2 Van der Helm HJ, Hische EAH. Beoordeling van de diagnostische waarde van klinisch-chemische onderzoekingen met behulp van het theorema van Bayes. Ned Tijdschr Geneeskd 1979b; 123: 1983-7.

3 Haynes RB. Hoe moeten medische tijdschriften worden gelezen? II. Het beoordelen van een diagnostische test. Ned Tijdschr Geneeskd 1983; 127: 2331-7.

4 Knottnerus JA, Volovics A. Het onderscheidend vermogen van diagnostische tests. Huisarts Wet 1989; 32: 338-46.

5 Begg CB. Biases in the assessment of diagnostic tests. Stat Med 1987; 6: 411-23.

6 Diamond GA. ROC steady; a Receiver Operating Characteristic curve that is invariant relative to selection bias. Med Decis Making 1987; 7: 238-43.

7 Hunink MGM, Richardson DK, Doubilet PM, Begg CB. Testing for fetal pulmonary maturity; ROC analysis involving covariates, verification bias and combination testing. Med Decis Making 1990; 10: 201-11.

8 Van der Schouw YT, Van der Veeken PMWC, Kok FJ, et al. Iron status in the acute phase and six weeks after myocardial infarction. Free Radical Biol Med 1990; 8: 47-53.

9 Miles LEM, Lipschitz DA, Dieber CP, Cook JD. Measurement of serum-ferritin by a two-side immunoradiometric assay. Anal Chem 1974; 61: 209-13.

10 Eskelinen S, Haikonen M, Räisänen S. Ferene-S as the chromogen for serum iron determination. Scand J Clin Invest 1983; 43: 453-55.

11 Goldman L. Quantative aspects of clinical reasoning. In: Jeffers JD, Scott EJ, Ramos-Englis M (eds). Harrison's principles of internal medicine. Tokyo: McGraw-Hill, 1987: 5-11.

12 Weinstein MC, Fineberg HV. Clinical decision analysis. London: WB Saunders Company, 1980.

13 Lusted LB. Decision-making studies in patient management. New Engl J Med 1971; 284: 416-24.

14 Swets JA. The relative operating characteristics in psychology. Science 1973; 182: 990-1000.

15 Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978; 8: 283-98.

16 Sackett DL, Haynes RB, Tugwell P. Clinical epidemiology; a basic science for clinical medicine. Boston: Little, Brown and Company, 1985.

17 Hanley JA, McNeill BJ. The meaning and use of the area under a receiver operating chacteristic (ROC) curve. Radiology 1982; 143: 29-36.

18 Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983; 148: 839-43.

19 Rombach JJ, Collette HJA, De Waard F, Slotboom JJ. Analysis of the diagnostic performance in breast cancer screening by relative operating characteristics. Cancer 1986; 58: 169-77.

20 Verbeek ALM, Van Dijck JAAM, Derkx RMJ, et al. Het bevolkingsonderzoek naar borstkanker; screeningsepidemiologische aspecten. Nijmegen: Landelijk Referentiecentrum Borstkankerscreening en sectie Epidemiologie KUN, 1990.

21 Miettinen OS. Theoretical epidemiology. Principles of occurrence research in medicine. New York: John Wiley & Sons, 1985.

22 Albert A, Harris EK. Multivariate interpretation of clinical laboratory data. New York: Marcel Dekker Inc, 1987. ■