De betekenis van de betrouwbaarheid en validiteit van de testmethode en de prevalentie van de ziekte bij de beslissing al of niet tot screening over te gaan

Enige bedenkingen bij een landelijke screening op cervixcarcinoom

DOOR DR. F. STURMANS TE DRUTEN*, DRS. W. F. M. DE HAES TE CAPELLE A/D IJSSEL** EN DRS. P. G. H. MULDER TE 's-GRAVENHAGE***

„Screening is net als minirokken en marihuana „in" aan het raken. Het heeft dan ook veel van de eigenschappen om iets „in" te doen raken. Het is bekend dat het zeer populair is in de Verenigde Staten; op radio- en televisieprogramma's wordt er met verve over gesproken evenals in de weekbladen en de gekleurde bijlagen hiervan. Er is alleen nog maar een aanmoedigend woord van een Beatle nodig en screening zou epidemisch zijn" (Cochrane en Elwood).

INLEIDING

Onder screening (bevolkingsonderzoek) wordt verstaan de uitvoering van een of meer gemakkelijk en snel toe te passen tests, onderzoekingen of andere procedures op een populatie teneinde de individuen te classificeren in groepen op basis van de waarschijnlijkheid dat deze een bepaalde ziekte of aandoening hebben *(figuur 1)*. Screeningtests verdelen ogenschijnlijk gezonden in personen die de ziekte waarschijnlijk wel en die de ziekte waarschijnlijk niet hebben.

Diagnostiek is het gedetailleerde onderzoek van individuen met het doel een definitieve toewijzing tot een ziekteclassificatie tot stand te brengen.

Een veel gehuldigde stelling is dat screening gelijk staat met vroege diagnostiek, dus zonder meer goed is en altijd toepassing verdient. Als de screening geschiedt met een valide methode (nagenoeg geen fout-positieven of fout-negatieven opleverend) èn betrouwbaar (= consistent = reproduceerbaar), lijkt deze stelling op het eerste oog te kunnen worden onderschreven. Doch ook als de screening geschiedt met een betrouwbare en valide methode dan blijkt de vraag, of screening in het onderhavige geval nuttig is, verder nog afhankelijk van de prevalentie van de ziekte waarop men de populatie screent.

BETROUWBAARHEID VAN EEN MEETMETHODE

Onder de betrouwbaarheid (= herhaalbaarheid = reproduceerbaarheid = repeatability) van een meetmethode wordt verstaan de mate waarin een methode hetzelfde resultaat oplevert bij hermeting van eenzelfde individu op twee of meer momenten. De meting van lichaamslengte is bijvoorbeeld vrij betrouwbaar; de meting van de bloeddruk is minder betrouwbaar zowel door de wisselvalligheid van de bloeddruk zelf als door subjectieve factoren bij de onderzoeker. Als het gaat om een ziekte geldt natuurlijk als voorwaarde dat het individu tussen de meetmomenten de ziekte niet heeft gekregen of van de ziekte is genezen. De betrouwbaarheid is afhankelijk van:

a de werkelijke of biologische variabiliteit in het te meten kenmerk, bijvoorbeeld bloeddruk varieert bij een en dezelfde persoon in de tijd zodat het ene moment de arbitraire scheidslijn tussen ziek en niet ziek wel en het andere moment niet wordt overschreden;

b de variabiliteit in de meting; deze is op haar beurt weer afhankelijk van:

1 de intrinsieke nauwkeurigheid van de meetmethode;

2 de bekwaamheid van de onderzoekers in het hanteren van de meetmethode en in het interpreteren en noteren van hun waarnemingen. Bij dit laatste is de volgende onderscheiding van belang:

– de intra-waarnemervariatie: het waarnemingsvermogen en oordeel van een en dezelfde onderzoeker kan van keer tot keer wisselen;

– de inter-waarnemervariatie: het verschil in waarnemingsvermogen en beoordelingskwaliteiten tussen twee of meer onderzoekers.

Stel dat men twee bepalingen doet bij een aantal personen. De resultaten dienen dan te worden weergegeven als in *schema 1*.

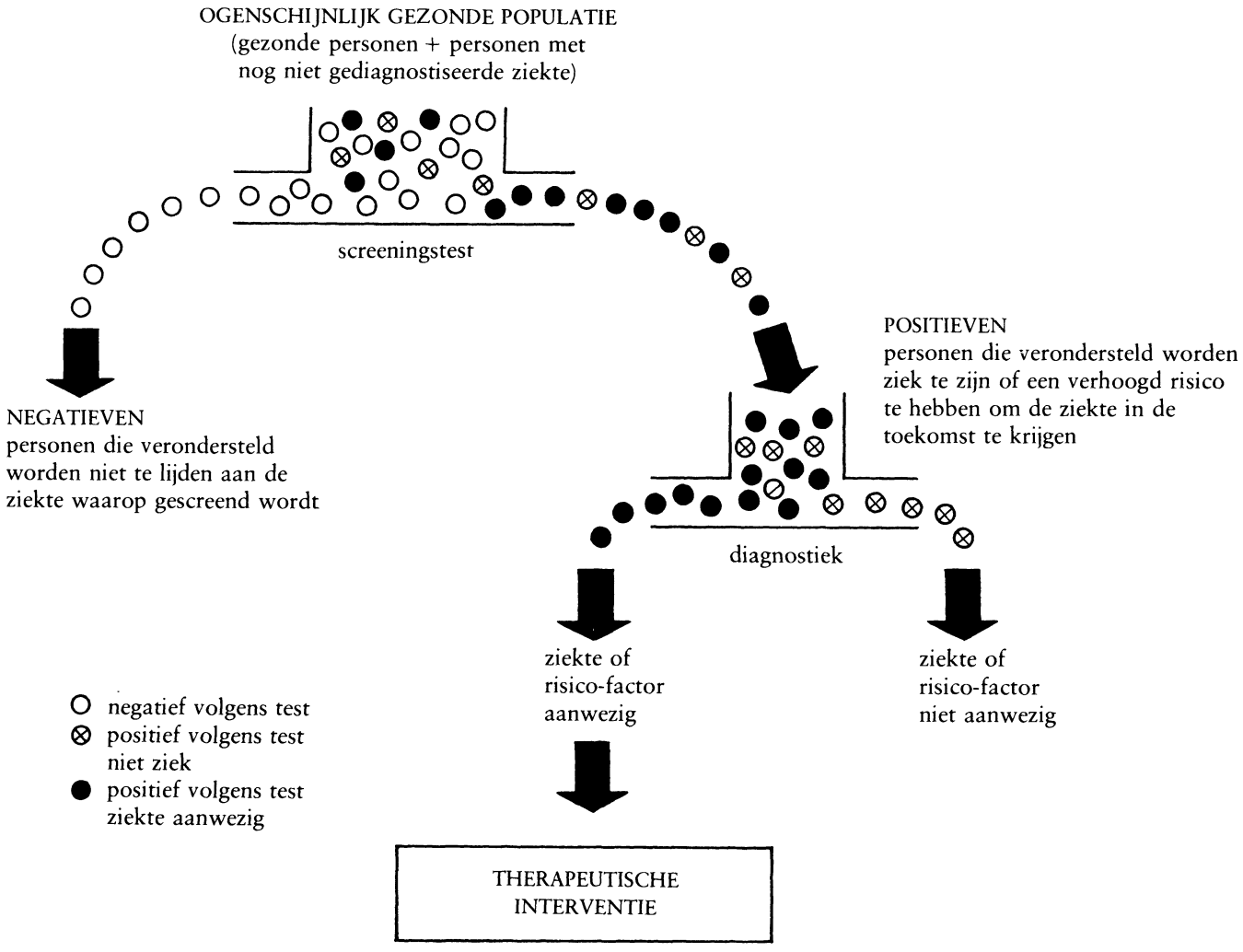
De herhaalbaarheid (= repeatability)

Samenvatting. In dit artikel wordt eerst ingegaan op de criteria ter beoordeling van de bruikbaarheid van een onderzoekmethode voor epidemiologisch onderzoek: de betrouwbaarheid, de sensitiviteit en de specificiteit van de methode. Vervolgens wordt aangetoond dat, ook al heeft men een voor epidemiologisch onderzoek aan de criteria voldoende onderzoekmethode, dit niet betekent dat hiermede tevens de voorwaarden zijn vervuld een screeningscampagne te beginnen. Hiertoe wordt nader ingegaan op de predictieve waarde van de test. Deze predictieve waarde blijkt in hoge mate afhankelijk te zijn van de prevalentie van de ziekte waarop men screent.

Voor vele ziekten wordt op dit moment screening bepleit. Voor bijvoorbeeld cervixcarcinoom blijkt de predictieve waarde van een positieve testuitslag ook bij een hoge mate van sensitiviteit (95%) en specificiteit (98%) slechts 16% te zijn. De conclusie moet dan ook zijn dat alleen high-risk-groepen zouden moeten worden gescreend. Hiervoor is de huisarts een onmisbare schakel.

* Hoofd, ** sociaal-psycholoog, *** econometrist afdeling Gezondheidsvoorlichting en -opvoeding van de Gemeentelijke Geneeskundige en Gezondheidsdienst te Rotterdam (Directeur Prof. Dr. L. Burema)

Figuur 1. Flow-diagram voor gang van zaken bij een bevolkingsonderzoek (naar Mausner en Bahn).



wordt uitgedrukt als het percentage

$$\frac{\text{aantal bij beide bepalingen positief}}{\text{aantal bij een of beide bepalingen positief}} = \frac{a}{a+b+c} \times 100.$$

In het ideale geval zou dit 100 moeten zijn, dit betekent dat b en c nul zijn (zie hiervoor verder *Sturmans* en *Mulder* 1976).

VALIDITEIT VAN EEN TEST

Onder validiteit of deugdelijkheid verstaat men de mate van overeenstemming tussen de uitslag c.q. de waarde welke de test oplevert en de mate van aanwezigheid van datgene wat de test wordt geacht te meten. Validiteit heeft dus betrekking op de vraag of de test werkelijk dat meet wat moet worden gemeten.

In de epidemiologie verstaat men verschillende dingen onder een test of meetinstrument. Men kan hieronder verstaan een gestandaardiseerde vragenlijst, bijvoorbeeld de Rose Questionnaire ter vaststelling van de vraag of iemand lijdt aan angina pectoris, een psychologische test bijvoorbeeld een intelligentie-test, een laboratorium-methode bijvoorbeeld strookjes-test ter vaststelling van bacteriurie, een fysisch-fysiologische methode bijvoorbeeld audiometrie of het maken van een electrocardiogram doch ook een klinisch-diagnostische bevinding bijvoorbeeld een palpabele lever of prostaat. In elk geval gaat het bij de bepaling van de validiteit om de vraag of een test in een kwantificeerbare verhouding staat tot dat wat de test eigenlijk meten moet, bijvoorbeeld tot een „diagnose". Ter vaststelling van de vraag of iemand een hartinfarct heeft wordt gebruik gemaakt van verhoogde S.G.O.T.- en S.G.P.T.-waarden doch in nog sterkere mate van de aanwezigheid van bepaalde afwijkingen in het electrocardiogram. Bij hoeveel van de werkelijke hartinfarct-patiënten zijn evenwel bedoelde electrocardiogramafwijkingen te zien? En anderzijds

Schema 1. Reproduceerbaarheid van een bepalingsmethode.

		1e bepaling	
		Positief	Negatief
2e bepaling	Positief	a	b
	Negatief	c	d

bij hoeveel van de onderzochten, die inderdaad geen hartinfarct (gehad) hebben, werden bedoelde afwijkingen in het electrocardiogram ook niet geconstateerd? Deze vragen zijn vragen naar de validiteit van de test.

De validiteit van een test hangt in zekere zin samen met de betrouwbaarheid (= reproduceerbaarheid). Een matige reproduceerbaarheid houdt noodzakelijkerwijs een matige validiteit in, daar slechts één antwoord het juiste kan zijn. Het omgekeerde geldt echter niet. Een methode kan namelijk consistent fout zijn: een fout afgestelde weegschaal geeft bij herhaalde meting bij eenzelfde persoon dezelfde, zij het foutieve uitslag (dus meet zeer betrouwbaar).

Er zijn diverse vormen van validering (*Sturmans en Mulder* 1976). In dit verband is van belang de zogenaamde diagnostische validering.

Schema 2. Een hypothetische verdeling van 500 personen die op CARA worden gescreend.

		Diagnose		
		CARA aanwezig	Geen CARA	Totaal
Test	Positief (+)	20	50	70
	Negatief (–)	5	425	430
	Totaal	25	475	500

Schema 3. Voor verklaring zie tekst.

	Ziekte aanwezig	Ziekte afwezig	
Screeningstest positief	a Terecht positief	b Fout positief	a + b
Screeningstest negatief	c Fout negatief	d Terecht negatief	c + d
Totaal	a + c	b + d	a + b + c + d
	Sensitiviteit $= \frac{a}{a+c} \times 100$	Specificiteit $= \frac{d}{b+d} \times 100$	
	Percentage fout-negatieven $= \frac{c}{a+c} \times 100$	Percentage fout-positieven $= \frac{b}{b+d} \times 100$	

DIAGNOSTISCHE VALIDERING

Bij de diagnostische validering gaat het om de vraag in hoeverre een diagnostische test, welke men vlug of bij vele mensen tegelijk kan afnemen, overeenstemt met de objectieve en tijdrovende diagnose welke men per individu moet uitvoeren. Verder gaat het er om welke kwantitatieve betrekking er bestaat tussen een screeningstest en de uiteindelijke diagnose en welke voorspelling een epidemiologisch bruikbare diagnostische test toestaat. Het begrip „epidemiologisch bruikbaar" heeft vooral betrekking op de praktische toepassing bij grote bevolkingsonderzoeken waarbij men veel individuen in korte tijd moet kunnen beoordelen. Hoewel in de epidemiologie validiteit op geen enkele wijze met diagnostische validiteit kan worden gelijkgesteld staat deze laatste toch in het middelpunt van de belangstelling.

Bijzonder actueel is de validiteitsbepaling ten aanzien van instrumenten die gebruikt worden in screeningsonderzoeken als vorm van vroege diagnostiek.

Onder validiteit wordt in dit concrete geval verstaan de mate waarin de resultaten van een vlug en massaal af te nemen test (bijvoorbeeld een screeningsprocedure) overeenkomen met de resultaten van een andere test (bijvoorbeeld een diagnostische procedure) die als nauwkeuriger of dichter bij de waarheid wordt beschouwd.

Validiteit van een instrument, dat de aan- of aanwezigheid van een ziekte of aandoening moet voorspellen, bestaat uit twee componenten: sensitiviteit en specificiteit. Voor het beoordelen van de zinvolheid van een bevolkingsonderzoek in de zin van screeningsonderzoek zijn deze componenten van het grootste belang.

SENSITIVITEIT EN SPECIFICITEIT

De sensitiviteit van een test is de gevoeligheid van de test: het vermogen van de test om alle personen met de betreffende ziekte te identificeren. Men kan deze gevoeligheid (sensitiviteit) bepalen door te berekenen hoeveel personen, die echt de ziekte hebben, (zoals uit een diagnose is gebleken) door de test worden ontdekt. Men drukt de sensitiviteit dan ook uit als een percentage: hoeveel procent van de werkelijk zieke gevallen wordt door de test ontdekt. Als bijvoorbeeld in een populatie van 500 mensen 25 gevallen van CARA aanwezig zijn (*schema 2*) en men kan er met een vragenlijst hiervan 20 ontdekken dan is de sensitiviteit:

$\frac{20}{25} \times 100 = 80$ procent.

Vanuit een algemeen schema, dat de resultaten van test en diagnose tegenover elkaar stelt (*schema 3*), is de sensitiviteit als volgt te berekenen:

$$\text{sensitiviteit} = \frac{\text{het aantal terecht positieven}}{\text{totaal aantal met ziekte}} \times 100 = \frac{a}{a+c} \times 100.$$

De specificiteit van een test is het vermogen van een test om uitsluitend personen met de betreffende ziekte te identificeren. Men gaat daarom bekijken hoeveel personen zonder ziekte ook door de test negatief worden bevonden. Men drukt dit eveneens uit als een percentage. In het vorige voorbeeld (*schema 2*) kan men zich voorstellen dat er van de 475 personen zonder CARA toch 50 positief scoren op de vragenlijst. Bij nader onderzoek kan dan blijken dat zij „fout-positief" zijn.

Een vragenlijst die op deze manier van de 475 gezonde mensen er 425 terecht als gezond beoordeelt, heeft een specificiteit van $\frac{425}{475} \times 100 = 89{,}5$ procent.

Met de algemene gegevens uit *schema 3*

wordt dit: specificiteit =

$$\frac{\text{het aantal terecht negatieven}}{\text{totaal aantal zonder ziekte}} \times 100 = \frac{d}{b+d} \times 100.$$

Sensitiviteit is dus een maat voor het percentage terecht positieve tests onder zieken, specificiteit voor het percentage terecht negatieve tests onder niet-zieken.

Er zijn ziekten waarvan de symptomen zo duidelijk zijn dat het mogelijk is tests te construeren, die zowel een hoge sensitiviteit als een hoge specificiteit hebben. Men zegt dan dat deze tests een groot „onderscheidend vermogen" hebben. Dit zijn de zogeheten pathognomonische bevindingen die praktisch nooit bij niet-zieken en praktisch altijd bij mensen met de betreffende ziekte voorkomen. Veelal gaat de eis van een hoge sensitiviteit van een test ten koste van de specificiteit en omgekeerd, zoals elders uitvoerig zal worden aangetoond (*Sturmans* en *Mulder* 1976). Het is met andere woorden zeer moeilijk om een test te ontwikkelen voor het ontdekken van ziekten in grote groepen, die zo valide is dat zij alle (of bijna alle) ziektegevallen opspoort zonder dat er tegelijk een aantal mensen zonder de ziekte eveneens als „verdacht op ziekte" worden aangewezen. Als het dan lukt met een test de meeste zieken te ontdekken, zal men vaak op de koop toe moeten nemen dat zich onder de niet-zieke personen relatief velen bevinden die toch een positieve testuitslag hebben. Omgekeerd, als de test zeer specifiek is (dat wil zeggen de meesten die niet ziek zijn hebben inderdaad een negatieve testuitslag) komt het vaak voor dat personen, die de ziekte wel hebben, eveneens door de mazen van het (screenings-)net vallen omdat zij een testresultaat behalen dat hen niet „verdacht" maakt.

Als men bij een bloedsuikeronderzoek een nuchtere waarde van 200 mg procent als kritieke grens aanhoudt, zullen onder de positieve testuitslagen zich nagenoeg uitsluitend diabetici bevinden. De grens is echter zo hoog, dat zeer vele diabetici niet worden ontdekt, omdat hun bloedsuikerwaarden lager liggen, ook al hebben zij diabetes. Men heeft in dit geval te doen met een hoge specificiteit doch met een lage sensitiviteit van de test.

PREDICTIEVE WAARDE VAN EEN TEST

Bij de vraag naar de sensitiviteit en specificiteit bekijkt men de resultaten van de vragenlijst vanuit de diagnostische gegevens. Men kan echter ook vanuit de vragenlijst-resultaten gaan bekijken wat de diagnostische gegevens kunnen opleveren. In het CARA-voorbeeld *(schema 2)* waren 70 testresultaten positief. Er waren slechts 20 personen die inderdaad CARA hadden. Iemand, die als „verdacht op ziekte" uit dit onderzoek komt, heeft dus een kans van slechts 2 op 7 dat hij inderdaad ziek is. Met de algemene gegevens van *schema 3*:

$$\frac{\text{het aantal terecht positieven}}{\text{alle test-positieven}} \times 100 = \frac{a}{a+b} \times 100 =$$

de predictieve waarde van een positieve uitslag.

Ten aanzien van een negatieve uitslag kan men een analoge redenering maken. In het CARA-voorbeeld waren 430 personen met een negatieve uitslag, 425 hadden inderdaad geen ziekte. Iemand met een negatieve uitslag heeft dus slechts een kans van 5 op 430 (dat wil zeggen ruim 1 op 100) dat hij toch CARA heeft. Met de gegevens van *schema 3:*

$$\frac{\text{het aantal terecht negatieven}}{\text{alle test-negatieven}} \times 100 = \frac{d}{c+d} \times 100 =$$

de predictieve waarde van een negatieve uitslag.

Onder de predictieve waarde van een test wordt dus verstaan het relatieve aantal werkelijk zieken onder de testpositieven respectievelijk het relatieve aantal werkelijk niet-zieken onder de test-negatieven. De predictieve waarde geeft dus aan in welke mate een juiste voorspelling van de ziekte-hebben of de ziekte niet-hebben vanuit de test mogelijk is.

Men kan ook het complement van deze uitslag bekijken, dat wil zeggen nagaan in welke mate een foutieve voorspelling wordt gegeven door de positieve of negatieve testuitslag. Met de algemene gegevens van *schema 3:*

$$\frac{\text{het aantal fout-positieven}}{\text{alle test-positieven}} \times 100 = \frac{b}{a+b} \times 100 =$$

de kans dat een positieve testuitslag fout-positief is.

$$\frac{\text{het aantal fout-negatieven}}{\text{alle test-negatieven}} \times 100 = \frac{c}{c+d} \times 100 =$$

de kans dat een negatieve testuitslag fout-negatief is.

DE ROL VAN DE PREVALENTIE VAN DE ZIEKTE

De zojuist gedefinieerde predictieve waarde van een positieve respectievelijk negatieve testuitslag hangt niet alleen af van de sensitiviteit en de specificiteit van de test maar ook van de prevalentie van die ziekte in de populatie.

A Neem als voorbeeld een test met 95 procent sensitiviteit en 98 procent specificiteit, bedoeld om een ziekte te detecteren met een prevalentie in de bevolking van 10 procent.

Zoals bekend, is de prevalentie het aantal ziektegevallen dat op een bepaald moment aanwezig is in een bepaalde bevolkingsgroep. In het CARA-voorbeeld zijn er 25 ziektegevallen op 500 individuen, de prevalentie is dus:

$$\frac{25}{500} \times 100 = 5 \text{ procent.}$$

Met de gegevens uit *schema 3:*

$$\text{prevalentie} = \frac{\text{aantal ziektegevallen}}{\text{de hele populatie}} \times 100 = \frac{a+c}{a+b+c+d} \times 100$$

Om de rol van de prevalentie van een ziekte voor de predictieve waarde van een testuitslag te verduidelijken zullen enkele voorbeelden worden besproken.

Wij gaan hierbij bewust uit van een zeer goede test. Met deze sensitiviteit van 95 procent en specificiteit van 98 procent wordt slechts 5 procent van alle zieken in de bevolking niet ontdekt en in de groep, die niet aan de ziekte lijdt, zijn slechts 2 op de 100 individuen toch testpositief. Uitgaande van een populatie van 10.000 personen ontstaat een situatie zoals in *schema 4* weergegeven. Men ziet duidelijk dat de sensitiviteit is:

$$\frac{950}{950+50} \times 100\% = \frac{950}{1000} \times 100\% = 95\%,$$

en dat de specificiteit is:

$$\frac{8820}{8820+180} \times 100\% = \frac{8820}{9000} \times 100\% = 98\%.$$

Men ziet ook duidelijk dat de preva-

lentie 1000 op 10.000 is, dus 10 procent.

De belangrijkste waarden zijn echter die ten aanzien van de predictiemogelijkheden van de test. De predictieve waarde van een positieve testuitslag is:

$$\frac{950}{950 + 180} \times 100\% = \frac{950}{1130} \times 100\% =$$

84,1%.

In 84,1 procent van de gevallen is een positieve uitslag dus ook een juiste uitslag: deze mensen hebben inderdaad de ziekte.

In de overige 15,9 procent is men niet ziek, ook al is de testuitslag positief. In het voorafgaande werd deze waarde de kans op een fout-positieve testuitslag genoemd:

$$\frac{180}{950 + 180} \times 100\% = \frac{180}{1130} \times 100\% =$$

15,9%.

De predictieve waarde van een negatieve testuitslag is:

$$\frac{8820}{50 + 8820} \times 100\% = \frac{8820}{8870} \times 100\% =$$

99,4 %

In 99,4 procent van de gevallen is iemand dus inderdaad niet ziek als de testuitslag negatief is.

De kans op een fout-negatieve testuitslag is:

$$\frac{50}{50 + 8820} \times 100\% = \frac{50}{8870} \times 100\% =$$

0,6%.

De kans is dus zeer klein dat men toch ziek is als de uitslag negatief is, slechts 1 op 200.

B In een tweede voorbeeld wordt uitgegaan van een testsituatie die even gunstig is als in het vorige geval: een zeer goede test met een sensitiviteit van 95 procent en een specificiteit van 98 procent. Men screent de bevolking met deze goede test echter op een ziekte die weinig voorkomt, namelijk met een prevalentie van 4 per 1000 of 0,4 procent. Een populatie van 10.000 personen levert dan de situatie op van *schema 5*. De gegeven waarden zijn gemakkelijk in de tabel te controleren:

sensitiviteit: $\frac{38}{40} \times 100\% = 95\%$.

specificiteit: $\frac{9760}{9960} \times 100\% = 98\%$.

prevalentie van de ziekte: 40 op 10.000 of 0,4 procent.

Schema 4. Voor verklaring zie tekst.

		Ziekte		
		Aanwezig	Afwezig	Totaal
Test	Positief	950	180	1130
	Negatief	50	8820	8870
	Totaal	1000	9000	10000

Hoe is het nu met de predictieve waarde van de testuitslagen? Een negatieve uitslag, in dit geval van een ziekte met een lage prevalentie, is vanzelfsprekend zeer predictief. Wij zeggen „vanzelfsprekend" omdat er slechts 4 zieken per 1000 mensen voorkomen. Als men dus zonder enig onderzoek van iedereen zegt dat hij „niet ziek" is, zal deze uitspraak in 99,6 van de gevallen juist zijn.

Met het gebruik van een goede test, zoals wij hier hebben aangenomen, moet men dus nog juister kunnen voorspellen. De gegevens zijn als volgt: predictieve waarde van een negatieve testuitslag:

$$\frac{9760}{9762} \times 100 = 99{,}98\%$$

dus praktisch 100%.

Kans dat een negatieve testuitslag fout-negatief is:

$$\frac{2}{9762} \times 100 = 0{,}02\%$$

dus praktisch 0%.

Iemand die een negatieve testuitslag behaalt, kan er dus nagenoeg zeker van zijn dat hij de ziekte niet heeft.

De predictieve waarde van een positieve testuitslag is echter slechts

$$\frac{38}{238} \times 100 = 16\%.$$

In woorden betekent dit, dat van degenen, die een positieve score op deze test behalen, er slechts 16 op de 100 ook inderdaad ziek zijn. De andere 84 personen zijn ten onrechte positief bevonden. In het geval van een ernstige ziekte heeft men dus 84 procent van de mensen die een positieve score halen onnodig ongerust gemaakt.

De gegevens van dit laatste voorbeeld krijgen een grote realiteitswaarde als men bedenkt dat 0,4 procent de geschatte prevalentie is van carcinoma-in-situ in Nederland voor vrouwen van 20 jaar en ouder. Als men aanneemt dat het beoordelen van de uitstrijkjes even juist kan gebeuren als in het tweede voorbeeld (sensitiviteit 95 en specificiteit 98 procent) zal toch nog steeds zo'n 85 procent van de positieve testuitslagen een „loos alarm" zijn en onnodige ongerustheid veroorzaken.

De moeilijkheid bij een bevolkingsonderzoek is dat men in vele gevallen niet over een test beschikt met een zo hoge sensitiviteit en specificiteit als in de voorbeelden. Het is nochtans van het grootste belang dat de specificiteit van een test zeer hoog is als de prevalentie van de op te sporen ziekte laag is. Wij zullen dit verduidelijken aan de hand van de gegevens uit de voorbeelden.

Als de prevalentie van de ziekte 10 procent is (voorbeeld 1) dan komt 90 procent van de mensen in de cellen b en d van *schema 3* terecht (= ziekte afwezig). Als dan de specificiteit 98 procent is, komt 2 procent van deze 90 procent in cel b terecht (in het eerste voorbeeld, *schema 4*, dus 180 personen van de 9000). Van de 1000 mensen die de ziekte wèl hebben komt als gevolg van de sensitiviteit van 95 procent, het grootste gedeelte, 950 personen, in cel a (= test positief) terecht.

De verhouding tussen de cellen a en b (950 en 180) is gunstig. Een positieve uitslag betekent meestal dat de persoon echt ziek is. Het percentage fout-positieve testuitslagen onder test-positieven is slechts 180 op 1130, dit is 15,9 procent.

Maar als de prevalentie van de ziekte slechts 0,4 procent is zoals in voorbeeld 2, wordt de situatie ongunstiger. Met een prevalentie van 0,4 procent komt 99,6 procent van de mensen in de cellen b en d terecht (= ziekte afwezig). In het tweede voorbeeld *(schema 5)* zijn dit 9960 personen. Met een specificiteit van 98 procent komen er hiervan 200 (= 2 procent) in cel b terecht. De 40 personen die de ziekte wèl hebben komen, als gevolg van de hoge sensitiviteit van 95 procent voor het grootste deel in cel a (= test positief) terecht, namelijk in 38 van de 40 gevallen. De grootte van de groep mensen die

de ziekte wèl heeft (a+c) is echter zo klein in verhouding tot de grootte van de groep mensen, die de ziekte niet heeft (b+d) dat de verhouding tussen a en b (38 en 200) nu zeer ongunstig is. Het percentage fout-positieve testuitslagen is nu 200 op 238, dit is 84 procent. Het grootste gedeelte van de positieve uitslagen komt dus voor bij personen die niet aan de ziekte lijden!

Men kan deze verhouding nauwelijks verbeteren door de sensitiviteit te verhogen (in het voorbeeld zou men dan moeten trachten van de 2 gevallen in c er nog 1 of alle 2 in a te krijgen). Men kan slechts trachten de specificiteit nog te verbeteren, dat wil zeggen trachten om de 200 gevallen uit b naar d te krijgen. Het gevolg zal zijn, jammer genoeg, dat er ook weer van de 38 gevallen in a enkele naar c gaan.

De conclusie moet zijn dat een zeer hoge waarde van de specificiteit (in dit geval 98 procent) nog onvoldoende is voor een bevolkingsonderzoek op een niet frequent voorkomende ziekte *(figuur 2)*. De meeste positieve uitslagen zouden bij nader onderzoek blijken niet op de ziekte te duiden. Dat 2 procent van de onderzochte groep een fout-positieve uitslag heeft lijkt niet veel, maar als men 100.000 mensen screent en de ziekte komt weinig voor, gaat het toch om zo'n 2.000 personen. Deze groep staat dan in een wanverhouding tot de kleine groep echt zieke personen in de groep met een positieve testuitslag. Het gevolg van deze situatie is dat zeer vele mensen onnodig ongerust worden gemaakt omdat zij na de screening worden opgeroepen voor een onderzoek door de arts, waarbij men verder moet bedenken dat de clinicus zich nauwelijks bewust is van hetgeen is betoogd.

C De situatie, zoals hiervoor geschetst, is in feite nog wat te rooskleurig. Daarom nog een derde voorbeeld.

Het is reëel te veronderstellen dat de opleiding van, in het geval van screening op cervixcarcinoom, de screensters gericht is op een hoge sensitiviteit (= alle ziektegevallen moeten zeker worden ontdekt). Het gevolg is echter, bijkans onherroepelijk, een lagere specificiteit (= er zullen heel wat gezonden toch een positieve uitslag halen). Als een meer reële schatting van de werkelijkheid kan men stellen: 100 procent sensitiviteit tegen 80 procent specificiteit. Als men met deze testkenmerken de gegevens van een ziekte met een prevalentie van 0,4 procent opniwue in schema brengt krijgt men *schema 6*. De predictieve waarde van een positieve testuitslag is dan:

$$\frac{40}{2032} \times 100 = 1{,}9 \text{ procent.}$$

Dit betekent dat van alle positieve testuitslagen slechts twee procent echt ziek zijn. De hoeveelheid fout-positieven is 98 procent!

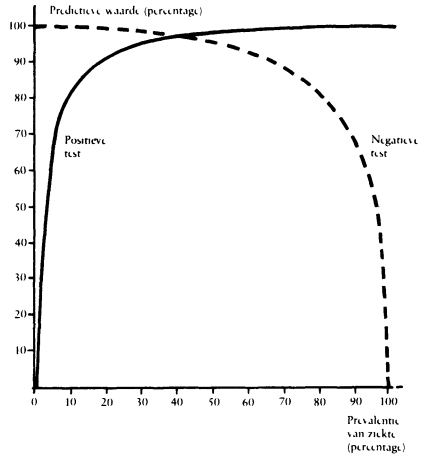
Men zou kunnen zeggen dat de screeningstest een populatie heeft afgezonderd waarin de prevalentie van de ziekte hoger is dan in de oorspronkelijke populatie. In het tweede voorbeeld *(schema 5)* is de prevalentie van de ziekte onder de testpositieven 16 procent geworden (dat wil zeggen 40 maal hoger dan in de oorspronkelijke populatie). In het tweede voorbeeld is de prevalentie van de ziekte onder de testpositieven 2 procent geworden (5 maal hoger dan in de oorspronkelijke populatie).

Laten wij ons optimistisch opstellen en aannemen, dat de eerste situatie (predictieve waarde van een positieve testuitslag = 16 procent) toch zou kunnen worden bereikt en op deze basis het vervolg van de procedure verder kritisch bekijken.

DE FASE NA DE SCREENING

Iedere clinicus zal natuurlijk onmiddellijk het bovenstaande als irrelevant van de hand wijzen en opwerpen dat hij niet over ijs van één nacht gaat en dat een positieve testuitslag altijd wordt gevolgd door nader onderzoek. In het geval van screening op baarmoederhalskanker door middel van een uitstrijkje volgens Papanicolaou betekent dit histologisch onderzoek van een biopsie van de cervix ter verificatie van de diagnose. Dit nader onderzoek heeft opnieuw zijn zwakten. Het betekent immers dat bij een specificiteit van de test van 98 procent, een sensitiviteit van 95 procent en een prevalentie van de ziekte van 0,4 procent de resultaten van 100 uit een bevolkingsonderzoek afkomstige positieve Pap-smears door de patholoog-anatoom in 84 gevallen als fout positief zouden moeten worden aangeduid. Dit doet waarschijnlijk geen enkele patholoog-anatoom. Hij weet namelijk dat hij alleen positieve preparaten ter beoordeling krijgt voorgelegd. Bovendien heeft hij een groot vertrouwen in de screenings-

Figuur 2. Relatie tussen prevalentie van ziekte en predictieve waarde bij een gegeven sensitiviteit van 95 procent en specificiteit van 98 procent.



Schema 5. Voor verklaring zie tekst.

		Ziekte		
		Aanwezig	Afwezig	Totaal
Test	Positief	38	200	238
	Negatief	2	9760	9762
	Totaal	40	9960	10000

Schema 6. Voor verklaring zie tekst.

		Ziekte		
		Aanwezig	Afwezig	Totaal
Test	Positief	40	1992	2032
	Negatief	–	7968	7968
	Totaal	40	9960	10000

procedure. Hij weet namelijk niet dat de kans op fout-positieven zo groot kan zijn. In zekere zin is hij dus bevooroordeeld: hij denkt dat hij in de meeste gevallen wel wat moet vinden. Nu is al vaak gebleken dat mensen, die op die manier „erop zijn gericht" iets te vinden, ook meestal wel wat vinden.

Rosenhan deed een experiment waarbij hij acht gezonde mensen zich liet aanmelden in veertien psychiatrische klinieken. Het resultaat was dat door de stafleden in totaal dertien maal de diagnose schizofrenie werd uitgesproken. Het omgekeerde gebeurde ook: hij sprak met een psychiatrische kliniek af dat zich in de volgende drie maanden af en toe pseudo-patiënten zouden aanmelden. In feite stuurde hij geen enkele pseudopatiënt. Van de 193 patiënten werd toen in 41 gevallen door ten minste één staflid gezegd dat de betrokkene een pseudopatiënt was. In 23 gevallen zei een psychiater dat de betrokkene een pseudopatiënt was.

Het zou al te kortzichtig zijn als men zou menen dat het beoordelen van histologische preparaten eenvoudiger is dan het onderkennen van geesteszieken in een groep gezonden of omgekeerd. Het feit dat de patholoog-anatoom er van uitgaat dat er aan de hem voorgelegde preparaten wel wat moet schelen „bevooroordeelt" hem dus sterk. Hij zal geneigd zijn veel mensen positief te bevinden. De patholoog-anatoom wordt echter door een andere factor nog meer bevooroordeeld.

Als goede clinicus is hij er namelijk van overtuigd (en dat is juist) dat het opgeven van een fout-negatieve uitslag „erger" is dan een fout-positieve uitslag. Het is waar dat het meestal erger is een ziekte niet te ontdekken dan iemand, die niet ziek is, toch te behandelen. Vraag is of dit ook voor cervix-carcinoom geldt. Enerzijds is immers bekend dat er een spontane genezing is van bepaalde gevallen van carcinoma in-situ, anderzijds is de enige „behandeling" van cervix-carcinoom vaak het wegnemen van de baarmoeder, wat toch een behoorlijke ingreep is als men in feite gezond is.

Vooral om deze instelling van de clinicus, maar mede beïnvloed door het algemene vooroordeel, zal de patholoog-anatoom van zichzelf een hoge sensitiviteit verlangen, hetgeen ten koste zal gaan van de specificiteit. In termen van *schema 3*: hij zal trachten in elk geval alle ziektegevallen te vinden en dus veel positieve testuitslagen afleveren (a en b hoog) met als gevolg dat hij ook heel wat niet zieken (uit d) bij de groep test-positieven (in b) zal rangschikken. Als de patholoog-anatoom een perfecte screener wil zijn, hoort hij dus in het aangehaalde voorbeeld van elke hem aangeboden 100 gevallen er 16 in a (ziek + uitslag onderzoek positief) en 84 in de (niet ziek + uitslag onderzoek negatief) te classificeren. Aangezien hij deze perfectie niet kan bereiken, dient gezocht naar mogelijkheden om deze situatie zo goed mogelijk te bereiken. Een eerste mogelijkheid daartoe zou er in kunnen bestaan de patholoog-anatoom te informeren dat er slechts ongeveer 20 procent positieve gevallen aanwezig zijn in het hem aangeboden materiaal. Deze aanpak kan erg nuttig zijn en kan de patholoog-anatoom aanmanen tot een zeer kritische beoordeling van het hem aangeboden materiaal. Het is echter ook mogelijk dat het tegengestelde effect zal gaan optreden, namelijk dat er te veel personen negatief worden beoordeeld.

Een betere procedure om de patholoog-anatoom te helpen in dit moeilijke selectiewerk zou als volgt kunnen worden opgezet. Tussen de positieve testuitslagen worden, in een onbekende en steeds wisselende verhouding, ook test-negatieven voorgelegd ter beoordeling. Op sommige dagen geeft men zelfs test-negatieven, op een ander moment uitsluitend test-positieven. De patholoog-anatoom zal dan uiterst alert en uiterst kritisch elk preparaat moeten bekijken om in elk geval de test-negatieven niet als positief te beschouwen. Op die manier zullen ook de test-positieven voldoende kritisch worden bekeken.

OPVOERING VAN DE PREDICTIEVE WAARDE VAN EEN TEST

De conclusie is dat bij de opzet van een screeningsonderzoek voor een ziekte met een lage prevalentie in eerste instantie moet worden gestreefd naar een screeningsinstrument met een hoge specificiteit, dat wil zeggen dat bij een sensitiviteit van minstens 95 procent een specificiteit nog hoger dan 98 procent wenselijk is. Men dient dus te zoeken naar een test met een zeer goede detectie van symptomen of een zeer valide combinatie van meer symptomen tot één uitslag.

Een andere mogelijkheid om de situatie te verbeteren is dat men tracht de te screenen groep zodanig te bepalen dat de prevalentie van de ziekte in de screeningsgroep hoger is dan de prevalentie in de totale populatie. Dit is mogelijk door alleen „high risk groepen" te gaan screenen. Voor de screening op cervix-carcinoom bijvoorbeeld zou men zich kunnen beperken tot de screening van vrouwen uit lagere sociaal-economische milieus met vele kinderen. Door zich te beperken tot deze groepen zal de predictieve waarde van een positieve testuitslag aanzienlijk worden vergroot, met andere woorden: men krijgt minder gevallen die ten onrechte ziek worden genoemd.

Een derde mogelijkheid is dat men een twee-traps-screening construeert. Met een eerste screening hanteert men dan een epidemiologisch eenvoudig te hanteren test, die met voldoende sensitiviteit en specificiteit (zoals in het tweede voorbeeld) praktisch alle ziektegevallen ontdekt plus de onvermijdelijke grote groep fout-positieven. Op deze beperkte groep van test-positieven, waarin de prevalentie toch aanzienlijk hoger is dan in de oorspronkelijke populatie, kan men dan een tweede screening uitvoeren met een meer gevoelig instrument, geschikt om in deze meer homogene groep toch nog een scheiding aan te brengen tussen fout-positieven en echte positieven. Alleen deze laatste groep zou dan door de artsen verder dienen te worden onderzocht.

Een vierde mogelijkheid gaat uit van de gedachte dat er geen tweede, meer gevoelig screeningsinstrument bestaat. In dat geval dient een arts (patholoog-anatoom) een tweede screening uit te voeren. Zoals in het tweede deel van „De fase na de screening" reeds werd beschreven, dient dan een situatie te worden gecreëerd waarin de arts voortdurend zeer alert en zeer kritisch alle preparaten moet bekijken, zich er van bewust zijnde dat de meeste test-positieven in feite fout-positieven zijn. Door deze

tweede fase zo uitermate kritisch op te zetten is enige garantie geboden dat niet al te veel vrouwen ten onrechte zullen worden behandeld.

Summary. In this article first of all attention is paid to criteria for judging the usefulness of a method for epidemiological research: the repeatability of the method, which is the rate at which a method gives the same result at remeasuring the same person, the sensitivity of the method and the specificity of the method. Secondly, it is demonstrated that an epidemiological research method meeting the above mentioned criteria sufficiently, does not necessarily imply fulfilment of the conditions to start a screening campaign. For this purpose the predictive value of the test method is considered. This predictive value appears to depend to a high degree on the prevalence of the disease to be screened.

At present for a number of diseases many pleas are made for screening. In the case of cervical cancer e.g. the predictive value of a positive test-result is estimated at 16% only, despite a high sensitivity (95%) and specificity (98%) of the test method. Therefore the conclusion should be to screen only high risk groups, for which the general practitioner is indispensable.

Cochrane, A. L. en P. C. Elwood. (1969), Med. Offr *121-122*, 53-57.

Mausner, J. S. en A. K. Bahn. Epidemiology, an introductory text. W. B. Saunders Company, Philadelphia, London, Toronto, 1974.

Rosenhan, D. L. On being sane in insane places, (1973), Science *179*, 250-258. (E. Laffrée besprak dit artikel uitvoerig in: (1973) Maandblad voor de Geestelijke Volksgezondheid, 28,— 273-279).

Sturmans, F. en P. G. H. Mulder. De beoordeling van de bruikbaarheid van een onderzoekmethode voor een epidemiologisch onderzoek. I. De betrouwbaarheid van de methode. II. De validiteit van de methode. Tijdschrift voor Sociale Geneeskunde (1976) (ter perse).

Sturmans, F. en P. G. H. Mulder. De onderlinge relatie tussen sensitiviteit en specificiteit (in voorbereiding).

Onderzoekingen rond het gezondheidscentrum Withuis IV: De maatschappelijk werker en zijn werk bezien door de cliënten

DOOR H. F. J. M. CREBOLDER, HUISARTS TE VENLO

Samenvatting. Verslag wordt gedaan van een onderdeel van een longitudinale studie in Venlo, via enquêtering, over de visie van de cliënten op de werkers en het werk van de eerstelijnsgezondheidszorg. Met name wordt aandacht besteed aan de invloed van het gezondheidscentrum Withuis op die visie. Het hier behandelde onderwerp betreft de maatschappelijk werker en zijn werk.

De respondenten werden verdeeld in Withuis-cliënten en cliënten behorende tot een controlegroep. Degenen die aangaven contact te hebben gehad met het maatschappelijk werk (ruim 10 procent) scoren hoger op de VOEG, zijn ontevredener over hun eigen gezondheid en hebben vaker kinderen dan de rest van de populatie. De zogenaamde ja-contacten werden in het Withuis significant vaker door de huisarts naar de maatschappelijk werker verwezen.

Samenwerking van huisarts en maatschappelijk werker wordt door de overgrote meerderheid van de respondenten nuttig geacht; in het Withuis in nog sterkere mate dan in de controlegroep. Het cliënt zijn van het Withuis en het in contact zijn geweest met het maatschappelijk werk kan een gang uit eigen beweging naar de maatschappelijk werker bevorderen. Het contact met het maatschappelijk werk beïnvloedt het oordeel over de bereikbaarheid van de maatschappelijk werker in positieve zin. Het cliënt zijn van het Withuis heeft daarop geen invloed.

Bij de taakbeoordeling van de maatschappelijk werker werden geen verschillen tussen ja- en neen-contacten of tussen Withuis en controlegroep gevonden. Met behulp van factor-analyse konden de dimensies „materiële" en „relationele hulpverlening" worden onderscheiden.

Toegespitst op concrete situaties bleken de cliënten van het Withuis de maatschappelijk werker in significant sterkere mate als hulpverlener te kiezen bij relationele problemen, dan de cliënten van de controlegroep.

INLEIDING

In deze bijdrage willen wij de uitkomsten bespreken van een onderzoek naar de opinie van (potentiële) cliënten over de taak en bereikbaarheid van de algemeen maatschappelijk werker (MW) en over zijn samenwerking met de huisarts; deze uitkomsten zijn onderdeel van een longitudinaal onderzoek dat reeds in 1972 is begonnen.

Voor zover wij weten is er zeer weinig studie verricht naar de beeldvorming van de cliënt ten aanzien van de MW als mogelijke hulpverlener. De enige ons bekende Nederlandse publikatie is: „Kennis en beeld van de sociale dienstverlening", in 1967 uitgegeven door het ministerie van Cultuur, Recreatie en Maatschappelijk werk. Het is een marktonderzoek – door middel van gesprekken met een representatief te achten steekproef van Nederlanders van 21 jaar en ouder – naar vraag en aanbod van de sociale dienst in het algemeen en het maatschappelijk werk in het bijzonder. Gegevens die betrekking hebben op kennis- en beeldfactoren en op attitudes werden verzameld. In 1971 werd een verkort herhalingsonderzoek verricht, dat in 1974 werd gepubliceerd (*Steenis-Perelaer*). Het in opdracht van C.R.M. gedane onderzoek had vooral een breed, beschrijvend en oriënterend karakter. Wij zijn in onze studie daarentegen uitgegaan van vooronderstellingen; ons onderzoek is meer van toetsende aard.

AANLEIDING

De MW behoort naast de huisarts en de wijkverpleegkundige tot het vaste trio dat aanwezig behoort te zijn in elk multidisciplinair samenwerkingsverband. Zijn aanwezigheid is zelfs een eis om in aanmerking te komen voor de „voorlo-