

**Diagnose en Voorspelling;
Enkele Bijdragen uit de Medische Statistiek**

M. Nijse

Sectie Persoonlijkeids- en Onderwijspsychologie; Rijksuniversiteit Groningen

Abstract

Assessment of the quality of diagnostic information in situations where both the predictor and the criterion are dichotomous usually is limited to the calculation of the (Bayesian) posterior probabilities of correct predictions (true positive and negative rates). Although the latter measures are of central importance, a broader view often is necessary as well as possible. By referring to medical, more specifically epidemiological research, it is demonstrated that a variety of concepts (e.g. sensitivity, specificity and attributable risk) and techniques (e.g. combining the information of two tests, constructing confidence intervals and testing hypotheses for posterior probabilities) are available, which fruitfully can be used in educational and psychological settings. The concepts and techniques are applied to the prediction of choice of VWO, one of the four types of Dutch secondary education, by means of a teacher rating and a test of general intelligence.

Key words: 2 x 2 tables, sensitivity, specificity, predictive values, attributable risk, confidence intervals, hypothesis testing.

The author is grateful to I. W. Molenaar, W. K. B. Hofstee and a reviewer for their comments on an earlier version of this paper. Thanks are due to T. A. B. Snijders for comments on two papers by Bennett. Requests for reprints should be sent to M. Nijse, Section of Personality and Educational Psychology, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands.

1. Inleiding

Het gebruik van diagnostische tests heeft vaak tot doel te voorspellen of een onderzochte persoon aan een criterium zal voldoen. Wanneer de test- en criteriumscores op een continue schaal liggen, wordt voor de evaluatie van de voorspellende waarde van de gebruikte test(s) gewoonlijk gebruik gemaakt van (meervoudige) regressievergelijkingen. Wanneer test(s) en/of criterium van nominaal of ordinaal niveau zijn, worden de uitkomsten meestal met log-lineaire modellen geanalyseerd.

In gevallen waarin test en criterium dichotoom zijn of gedichotomiseerd kunnen worden, worden de gegevens gewoonlijk in een 2×2 -tabel samengevat. Voor de bepaling van de voorspellende waarde van de gebruikte test wordt de Bayesiaanse of aposteriorikans op een goede voorspelling berekend. Meer specifiek: men bepaalt de kans op een positief criterium bij een positieve diagnose en de kans op een negatief criterium bij een negatieve diagnose. Omdat dergelijke berekeningen goed aansluiten bij veel praktische vraagstellingen en kansen een ondubbelzinnige en voor buitenstaanders inzichtelijke interpretatie hebben, is het gebruik van 2×2 -tabellen tamelijk populair.

In sommige onderzoeken hoeft met de bepaling van aposteriorikansen niet te worden volstaan. Het is bijvoorbeeld mogelijk voor deze kansen betrouwbaarheidsintervallen op te stellen of, indien twee tests zijn gebruikt, te toetsen of de voorspellende waarden van deze tests significant van elkaar verschillen. In andere gevallen heeft men behoefte aan een techniek waarmee de voorspellende waarden van twee of meer dan twee tests kunnen worden gecombineerd.

Ontwikkelingen in de medische statistiek hebben het mogelijk gemaakt op deze en andere vragen een antwoord te geven. Sommige ontwikkelingen zijn van recente datum, andere hebben geleid tot begrippen en technieken die reeds lang deel uitmaken van het gevestigde medische arsenaal. In dit artikel wordt getracht een overzicht te bieden van een aantal belangrijke begrippen en technieken die in de sociaal-wetenschappelijke diagnostiek vrijwel onbekend zijn. Het overzicht is beperkt tot onderzoeken waarin zowel de test als het criterium dichotoom zijn of gedichotomiseerd kunnen worden. In het laatste geval zijn ook andere dan de hier besproken technieken mogelijk (bijv. logistische regressie of correlatiecoëfficiënten), maar deze passen niet bij frequent voorkomende praktijksituaties waarin een dichotomisering op de test (bijv. onvoldoende - voldoende) en het criterium

(bijv. gezakt - geslaagd) beter aansluit bij vraagstellingen waarin interpretatie van kansen centraal staat.

In de medisch-statistische literatuur wordt voor niet-experimenteel onderzoek een onderscheid gemaakt in drie onderzoeksoptellingen (Fleiss, 1981): naturalistisch of observatief (cross-sectional) onderzoek, prospectief (cohort) onderzoek en retrospectief (case-control) onderzoek. Sommige statistische maten hebben voor ieder van de genoemde optellingen een andere operationalisatie. In dit artikel wordt uitgegaan van observationeel onderzoek, omdat dit type onderzoek in differentieel-psychologische toepassingen het meest voorkomt.

De besproken begrippen en technieken worden geïllustreerd aan de hand van gegevens met betrekking tot de voorspelling van de schoolkeuze van 524 leerlingen van klas 5 (tegenwoordig groep 7) na de lagere school (Nijssse, 1992). Omdat de leerlingen afkomstig zijn uit het Noorden van het land zijn de resultaten niet landelijk representatief. Men moet niet alleen rekening houden met de mogelijkheid dat in andere regio's min of meer verschillende adviespatronen bestaan, maar ook met het gegeven dat de grootten van apriori- en aposteriori-kansen niet onafhankelijk van elkaar zijn.

De klasseleerkracht gaf aan het einde van het cursusjaar een oordeel over het schooltype voortgezet onderwijs (LBO, MAVO, HAVO of VWO) dat hij voor een leerling het meest geschikt achtte. Ongeveer een half jaar later werd aan de leerlingen de GALO-intelligentietest (Kema & Kouwer, 1958) afgenomen. Op grond van het behaalde IQ werden leerlingen als LBO-, MAVO-, HAVO- of VWO-leerling geclassificeerd wanneer zij voor deze schooltypen een volgens de handleiding "redelijke" kans op succes hadden. In dit artikel worden alleen de gegevens van Schooloordeel en GALO-test in relatie tot het gekozen schooltype na de brugklas, gedichotomiseerd als VWO resp. niet-VWO, gebruikt. Voor een meer volledig overzicht wordt verwezen naar Nijssse (1994). De gegevens zijn samengevat in Tabel 1; de benodigde 2 x 2 -tabellen vindt men in Tabel 2.

Alvorens op de gegevens in te gaan, worden in de volgende paragraaf eerst enkele basisbegrippen besproken. In Paragraaf 3 komen enkele aspecten m.b.t. het gebruik van aposteriori-kansen aan de orde. Het attributie-risico, een maat voor de bepaling van de sterkte van het oorzakelijk verband tussen predictor en criterium, is onderwerp van Paragraaf 4. Enkele problemen bij het gebruik van twee tests worden in Paragraaf 5 behandeld. In Paragraaf 6 wordt een slotbeschouwing en samenvatting van de resultaten gegeven.

Tabel 1. Schooloordeel (O) en GALO-test (G) in relatie tot VWO- (V) resp. niet-VWO-keuze (V).

Diagnose		Uitkomst		
<u>O</u>	<u>G</u>	<u>V</u>	<u>V</u>	Tot.
+	+	43	3	46
+	-	5	5	10
-	+	44	72	116
-	-	13	339	352
		-----	-----	-----
Tot.		105	419	524

Tabel 2. Samenvatting van de gegevens van Tabel 1 in twee 2 x 2 -tabellen.

A. Schooloordeel				B. GALO-test			
	<u>V</u>	<u>V</u>	Tot.		<u>V</u>	<u>V</u>	Tot.
<u>O</u>	48	8	56	<u>G</u>	87	75	162
<u>O</u>	57	411	468	<u>G</u>	18	344	362
Tot.				Tot.	105	419	524

2. Basisbegrippen

Een schema voor de weergave van onderzoeksgegevens bij een dichotome diagnose en een dichotoom criterium is weergegeven in Tabel 3. Hierin is p_{11} de (steekproef-)proportie individuen die zowel een positieve diagnose als een positief criterium heeft; p_{12} is de proportie met een positieve diagnose maar een negatief criterium, p_{21} de proportie met een negatieve diagnose maar een positief criterium en p_{22} de proportie met een negatieve diagnose en een negatief criterium. De basisverhouding (BV) is $p_{.1}$ en de selectieverhou-

Tabel 3. Tabel voor uitkomsten van een onderzoek met een dichotome diagnose en een dichotoom criterium.

		Uitkomst (criterium)		
		<u>B</u>	<u>$\bar{B}$</u>	Tot.
Diagnose	<u>A</u>	p_{11}	p_{12}	$p_{1.}$
(voorspelling)	<u>$\bar{A}$</u>	p_{21}	p_{22}	$p_{2.}$
		----	----	----
	Tot.	$p_{.1}$	$p_{.2}$	$p_{..}$

ding (SV) $p_{1.}$. Wanneer p_{11} tot en met p_{22} worden gestandaardiseerd door ze te delen door de bijbehorende marginale rij-proporties worden de volgende voorwaardelijke kansen verkregen: terecht-positieven-verhouding (TPV), onterecht-positieven-verhouding (OPV), onterecht-negatieven-verhouding (ONV) en terecht-negatieven-verhouding (TNV). In de medische diagnostiek worden p_{11} tot en met p_{22} niet alleen horizontaal, maar ook verticaal genormeerd. De sensitiviteit (Se) is de kans dat een persoon met een positief criterium een positieve test heeft: $p_{11}/p_{1.} = p(\underline{A} | \underline{B})$. De specificiteit (Sp) is de kans dat een persoon met een negatief criterium een negatieve test heeft: $p_{22}/p_{2.} = p(\underline{\bar{A}} | \underline{\bar{B}})$. Deze begrippen werden o.a. gebruikt door Yerushalmy (1947).

Tabel 4. Voorwaardelijke kansen en begrippen bij normering voor $p(\underline{A})$ resp. $p(\underline{\bar{A}})$.

		Uitkomst		
		<u>B</u>	<u>$\bar{B}$</u>	Tot.
	<u>A</u>	$p(\underline{B} \underline{A})$	$p(\underline{\bar{B}} \underline{A})$	1
		(TPV)	(OPV)	
Diagnose	<u>$\bar{A}$</u>	$p(\underline{B} \underline{\bar{A}})$	$p(\underline{\bar{B}} \underline{\bar{A}})$	1
		(ONV)	(TNV)	
		-----	-----	-----
	Tot.	—	—	—

Tabel 5. Voorwaardelijke kansen en begrippen bij normering voor $p(\underline{B})$ resp. $p(\overline{B})$.

	Uitkomst		Tot.
	$\underline{B}$	$\overline{B}$	
$\underline{A}$	$p(\underline{A} \underline{B})$ (Se)	$p(\underline{A} \overline{B})$ (1 - Sp)	-
Diagnose			
$\overline{A}$	$p(\overline{A} \underline{B})$ (1 - Se)	$p(\overline{A} \overline{B})$ (Sp)	-
-----	-----	-----	-----
Tot.	1	1	-

Voor de overzichtelijkheid worden de genoemde voorwaardelijke kansen en bijbehorende begrippen in Tabel 4 en Tabel 5 vermeld. (Voor het uitvoeren van berekeningen geeft men meestal de voorkeur aan tabellen met frequenties (N_{ij}) in plaats van proporties).

Sensitiviteit en specificiteit zijn, evenals de terecht-positieven- en terecht-negatieven-verhouding, belangrijke begrippen die nuttige informatie over de gebruikswaarde van tests verschaffen. Wanneer het gestandaardiseerde tests betreft, worden sensitiviteit en specificiteit meestal als steekproef-onafhankelijke maten gebruikt. De terecht-positieven- en terecht-negatieven-verhouding worden vaak positive predictive value resp. negative predictive value genoemd (Vecchio, 1966).

Wanneer de sensitiviteit en/of de specificiteit niet volmaakt (< 1) zijn, treden foute classificaties op; de terecht-positieven- en/of terecht-negatieven-verhouding zijn dan evenmin volmaakt. Youden (1950) toonde aan dat

$$\underline{J} = (\underline{Se} + \underline{Sp} - 1) > 0 \quad (1)$$

een voorwaarde is voor zinvol testgebruik. (De begrippen sensitiviteit en specificiteit werden door Youden overigens niet gebruikt). Een test waarvoor $\underline{J} = 0$ noemde hij waardeloos: een dergelijke test geeft geen informatie over de ware status ($\underline{B}$ of $\overline{B}$) van een onderzochte persoon. De classificatie is in zo'n geval niet beter dan wanneer deze door het toeval op grond van basis- en selectieverhouding zou hebben plaatsgevonden. Wat op hetzelfde neerkomt: wanneer voor alle personen met testuitslag $\underline{A}$ ($\overline{A}$) criterium $\underline{B}$ ($\overline{B}$)

wordt voorspeld, is de proportie goede voorspellingen gelijk aan de basisverhouding (het complement van de basisverhouding).

Uit de gegevens van Tabel 1 en 2 blijkt dat 105 van de 524 leerlingen na de brugklas naar het VWO gaan; de basisverhouding is dus .20. In vergelijking met deze waarde is de selectieverhouding van het Schooloordeel (Tabel 2A) te klein (.11) en van de GALO-test (Tabel 2B) te groot (.31). Vooral als gevolg van de scherpe selectieverhouding van het Schooloordeel is de sensitiviteit gering: meer dan de helft van de VWO-leerlingen wordt niet als zodanig onderkend. De terecht-positieven-verhouding is daarentegen groot (.86). De GALO-test heeft een grotere sensitiviteit (.83), maar een kleinere terecht-positieven-verhouding (.54) dan het Schooloordeel. De absolute verschillen tussen de specificiteiten van Schooloordeel en GALO-test (.98 resp. .82) en de terecht-negatieven-verhoudingen (.88 resp. .95) zijn minder groot.

Uit de eis $\underline{J} = \underline{Se} + \underline{Sp} - 1 > 0$ kan worden afgeleid $\underline{Se} > 1 - \underline{Sp}$. In termen van voorwaardelijke kansen: $p(\underline{A} | \underline{B}) > p(\underline{A} | \overline{\underline{B}})$. Bewezen kan worden dat in dat geval ook geldt: $p(\underline{B} | \underline{A}) > p(\underline{B} | \overline{\underline{A}})$. In woorden: als de sensitiviteit groter is dan het complement van de specificiteit, is de terecht-positieven-verhouding groter dan de onterecht-negatieven-verhouding. Op analoge wijze kan uit $\underline{J} > 0$ worden afgeleid $\underline{Sp} > 1 - \underline{Se}$, waaruit volgt dat de terecht-negatieven-verhouding groter is dan de onterecht-positieven-verhouding. De maat $\underline{J}$ speelt in vele formules een belangrijke rol.

Youden wees erop dat gebruik van $\underline{J}$ veronderstelt dat sensitiviteit en specificiteit even belangrijk zijn. In de praktijk is dit niet altijd het geval. In medisch onderzoek kan een lage sensitiviteit (relatief veel zieken worden ten onrechte voor gezond aangezien) veel ernstiger gevolgen hebben dan een lage specificiteit. Ook in het geval van de keuze van een geschikt schooltype kan een lage sensitiviteit schadelijk zijn: de keuze van een te laag schooltype is in het licht van de geringe mogelijkheden voor tussentijds "opstromen" een ernstige fout.

Bross (1954) toonde als eerste aan dat, onafhankelijk van de steekproefgrootte, de steekproef-selectieverhouding, die door de onderzoek(st)er verondersteld wordt gelijk te zijn aan de steekproef-basisverhouding, geen zuivere schatter van de populatie-basisverhouding is als de sensitiviteit en/of specificiteit niet volmaakt zijn. Quade, Lachenbruch, Whaley, McClish en Haley (1980) werkten dit resultaat verder uit; zij presenteerden onder meer een formule voor de standaardfout van de onzuiver geschatte basisverhouding en toonden aan dat

$$[\underline{SV} - (1 - \underline{Sp})]/I \quad (2)$$

een zuivere schatter van de populatie-basisverhouding is wanneer de sensitiviteit en specificiteit bekend verondersteld worden.

Omdat de sensitiviteit en specificiteit binomiale variabelen zijn, kunnen betrouwbaarheidsintervallen worden opgesteld met behulp van een logaritmische transformatie voor $\underline{O}_1 = \underline{Se}/(1 - \underline{Se})$ resp. $\underline{O}_2 = \underline{Sp}/(1 - \underline{Sp})$. Na terugtransformatie berekent men:

$${}_p\underline{Se} = (1 + \underline{O}_1^{-1})^{-1} \text{ en } {}_p\underline{Sp} = (1 + \underline{O}_2^{-1})^{-1} . \quad (3)$$

De onder- en bovengrens van het .95 -betrouwbaarheidsinterval voor de sensitiviteit van het Schooloordeel zijn .36 resp. .55; voor de GALO-test .74 resp. .89. Voor de specificiteiten zijn de waarden .96 en .99 (Schooloordeel) en .78 en .85 (GALO-test).

3. Aposteriori-kansen

Zoals bekend, wordt de kwaliteit van een diagnostische test meestal afgemeten aan het verschil tussen de kans op een juiste voorspelling in het licht van de testuitslag (de aposteriori-kans $p(\underline{B} | \underline{A})$) en de kans op een juiste voorspelling zonder testuitslag (de apriori-kans $p(\underline{B})$). Met de bekende formule van Bayes kan de aposteriori-kans als volgt worden berekend:

$$p(\underline{B} | \underline{A}) = \frac{p(\underline{B}) \cdot p(\underline{A} | \underline{B})}{p(\underline{A})} = \frac{p(\underline{A} | \underline{B})p(\underline{B})}{p(\underline{A} | \underline{B})p(\underline{B}) + p(\underline{A} | \underline{\bar{B}})p(\underline{\bar{B}})} . \quad (4)$$

Met behulp van de begrippen terecht-positieven-verhouding, basisverhouding, selectieverhouding, sensitiviteit en specificiteit kan (4) worden geschreven als:

$$\underline{TPV} = \underline{Se} \cdot \frac{\underline{BV}}{\underline{SV}} = \frac{\underline{Se.BV}}{\underline{Se.BV} + (1 - \underline{Sp})(1 - \underline{BV})} . \quad (5)$$

Voor $p(\bar{B} | \bar{A})$, zijn de analoge formules:

$$p(\bar{B} | \bar{A}) = p(\bar{A} | \bar{B}) \frac{p(\bar{B})}{p(\bar{A})} = \frac{p(\bar{A} | \bar{B})p(\bar{B})}{p(\bar{A} | \bar{B})p(\bar{B}) + p(\bar{A} | B)p(B)} \quad (6)$$

en

$$\underline{TNV} = \underline{Sp} \frac{(1 - \underline{BV})}{(1 - \underline{SV})} = \frac{\underline{Sp}(1 - \underline{BV})}{\underline{Sp}(1 - \underline{BV}) + (1 - \underline{Se})\underline{BV}} \quad (7)$$

Bij een positieve test wordt soms de kans op een juiste voorspelling in verhouding tot de kans op een onjuiste voorspelling uitgedrukt. Men berekent dan de aposteriori-kansverhouding $\underline{KV}_1 = p(B | A) / [1 - p(B | A)] = p(B | A) / p(\bar{B} | A)$:

$$\underline{KV}_1 = \frac{p(B | A)}{p(\bar{B} | A)} = \frac{p(A | B)}{p(A | \bar{B})} \times \frac{p(B)}{p(\bar{B})} \quad (8)$$

(8) kan worden herschreven als:

$$\underline{KV}_1 = \frac{\underline{TPV}}{\underline{OPV}} = \frac{\underline{Se}}{1 - \underline{Sp}} \times \frac{\underline{BV}}{1 - \underline{BV}} \quad (9)$$

$\underline{Se} / (1 - \underline{Sp})$ wordt meestal de likelihood ratio genoemd, maar staat ook bekend als de diagnostische verhouding (Baron, 1988, p. 190) of diagnosticiteit (Evans, 1991, p. 60). Hoe groter deze waarde is, hoe groter de verhouding $\underline{TPV} / \underline{OPV}$ bij een gegeven basisverhouding. $\underline{BV} / (1 - \underline{BV})$ staat bekend als de apriori-kansverhouding. Vanzelfsprekend kan men analoog aan (8) en (9) de aposteriori-kansverhouding voor $\underline{KV}_2 = p(\bar{B} | \bar{A}) / p(B | \bar{A}) = \underline{TNV} / \underline{ONV}$ opstellen.

De toename in de kans op een juiste diagnose dankzij de verkregen testinformatie werd nader bestudeerd door Connell en Koepsell (1985). Zij ontwikkelden o.a. formules voor de

winst in zekerheid voor een positieve test:

$$\frac{TPV - BV}{\frac{BV(1 - BV)J}{Se.BV + (1 - Sp)(1 - BV)}} \quad (10)$$

Merk op dat uit deze formule opnieuw duidelijk blijkt dat het afnemen van een test zinloos is wanneer $J = 0$. Connell en Koepsell toonden aan dat in het geval $J > 0$, de toename van de winst voor een positieve test groter is bij een toename van de specificiteit dan bij een zelfde toename van de sensitiviteit. Verder is de toename van de winst maximaal als

$$BV = \{[Se(1 - Sp)]^{1/2} + (Sp - 1)\} / J \quad (11)$$

Belangrijk is te vermelden dat het maximaliseren van de winst van een positieve test ten koste gaat van het maximaliseren van de winst van een negatieve test. Voor de toename van de winst van een negatieve test gelden dezelfde uitkomsten "in spiegelbeeld".

Soms wordt de winst uitgedrukt in termen die aan de informatie-theorie ontleend zijn. Asch, Patton en Hershey (1990) leverden hierop kritiek en stelden een verandering voor die erop neerkomt dat de informatie van een negatieve test niet negatief, maar positief geïnterpreteerd wordt. De totale hoeveelheid verwachte informatie $E(I)$ kan worden uitgedrukt als

$$E(I) = SV(TPV - BV) + (1 - SV)(BV - ONV) = 2BV(1 - BV)J \quad (12)$$

Diamond (1991) bestreed de ideeën van Asch e.a. De verdediging van Asch, Patton en Hershey (1991) is echter overtuigend.

Recent werd door Monsour, Evans en Kupper (1991) een methode ontwikkeld voor het opstellen van betrouwbaarheidsintervallen voor $P(B|A)$, de populatiewaarde van de aposteriori-kans op een juiste positieve diagnose. Zij baseerden hun methode op resultaten van Diamond en Forrester (1983) en Heckering (1988). Als uitgangspunt diende de vooronderstelling dat de drie ingrediënten van $P(B|A)$, sensitiviteit, specificiteit en basisverhouding, onafhankelijke, binomiaal-verdeelde variabelen zijn. Monsour e.a.

berekenen eerst een betrouwbaarheidsinterval voor $\mu_{KV_1}^{-1}$, de populatiewaarde van KV_1^{-1} , waarbij een logaritmische transformatie werd toegepast om de scheefheid van de steekproevenverdeling te reduceren. Door terugtransformatie en gebruikmaking van de gelijkheid

$$P(\underline{B} | \underline{A}) = [1 + \mu_{KV_1}^{-1}]^{-1} \quad (13)$$

wordt een $(1 - \alpha)$ -betrouwbaarheidsinterval voor $P(\underline{B} | \underline{A})$ verkregen. Voor een rechtstreekse berekening gebruikte men:

$$P(\underline{B} | \underline{A}) = \frac{1}{[1 + \mu_{KV_1}^{-1} \exp \pm z_{1-\alpha/2} \hat{V}^{1/2}]} \quad (14)$$

In (14) is z de z -waarde van de standaardnormale verdeling bij $1 - \alpha/2$; de geschatte variantie van $\ln(KV_1)^{-1}$ is

$$\hat{V} = \frac{1 - \underline{Se}}{\underline{N} \cdot \underline{Se}} + \frac{\underline{Sp}}{\underline{N} \cdot (1 - \underline{Sp})} + \frac{1}{\underline{N} \cdot \underline{BV}(1 - \underline{BV})} \quad (15)$$

Formules voor de berekening van betrouwbaarheidsintervallen voor $\mu_{KV_2}^{-1}$ en $P(\underline{B} | \underline{A})$ zijn analoog aan (14) te ontwikkelen.

De betrouwbaarheidsintervallen voor de aposteriori-kansen zijn vermeld in Tabel 6.

Tabel 6. .95-betrouwbaarheidsintervallen voor aposteriori-kansen voor Schooloordeel en GALO-test. (Puntschattingen tussen haakjes).

	Schooloordeel	GALO-test
$P(\underline{B} \underline{A})$	.75 - .92 (.86)	.46 - .61 (.54)
$P(\underline{B} \underline{A})$	.84 - .90 (.88)	.92 - .97 (.95)

4. Attributie-risico

In de medische wetenschap heeft men regelmatig te maken met het verschijnsel dat een bepaalde ziekte B zich niet alleen voordoet bij personen met een schadelijke levensgevoorte of werkomstandigheid A, maar ook bij personen bij wie A ontbreekt. Het verschil tussen beide gebeurtenissen is te schrijven als $p(\underline{B} | \underline{A}) - p(\underline{B} | \overline{A})$. De omgekeerde situatie is eveneens interessant: de ziekte B kan afwezig zijn als A afwezig, maar ook als A aanwezig is. De verklaring voor deze uitkomsten luidt meestal dat (de afwezigheid van) de ziekte door meer dan één factor kan worden veroorzaakt. Door Fleiss (1981, p. 76) werden de voorwaardelijke kansen $p(\underline{B} | \underline{A})$ en $p(\underline{B} | \overline{A})$ gewogen met $p(\underline{A})$ en het gewogen verschil gedeeld door $p(\underline{B})$; de aldus verkregen maat staat bekend als (Levin's) attributie-risico (AR):

$$\underline{AR} = \frac{p(\underline{B} | \underline{A})p(\underline{A}) - p(\underline{B} | \overline{A})p(\underline{A})}{p(\underline{B})} \quad (16)$$

AR kan worden geïnterpreteerd als de proportie zieke personen die aan de invloed van A kan worden toegeschreven. Wanneer teller en noemer door $p(\underline{A})$ worden gedeeld en kansen worden vervangen door begrippen, kan (16) worden herschreven als:

$$\underline{AR} = \frac{\underline{TPV} - \underline{ONV}}{\underline{BV/SV}} \quad (17)$$

Voor een niet-waardeloze test ($J > 0$) geldt: $0 < \underline{AR} \leq 1$.

Op dit punt is het van belang op te merken dat onlangs een nieuwe associatiemaat voor 2×2 -tabellen is ontwikkeld, die conceptueel verschilt van de maat AR, maar algebraïsch identiek is (Nijse, 1992). Copas en Loeber (1990) noemden deze nieuwe maat Relative Improvement Over Chance (RIOC); de maat kan worden gedefinieerd als de voor kans gecorrigeerde waarde p_{11} , gestandaardiseerd voor de maximaal mogelijke waarde van p_{11} (p_{11}), eveneens gecorrigeerd voor kans:

$$\text{RIOC} = \frac{P_{11} - P_{1.} P_{.1}}{P_{.1} - P_{1.} P_{.1}} \quad (18)$$

Een belangrijk verschil tussen beide maten is dat de interpretatie van AR altijd een causaal verband tussen A en B veronderstelt (Fleiss, 1981, pp. 75-76).

In de literatuur wordt veel aandacht besteed aan het schatten van AR bij de aanwezigheid van meerdere schadelijke factoren die geassocieerd zijn met de onderzochte factor en eveneens invloed hebben op de criterium-uitkomst (interaction en confounding). Walter (1978) wees erop dat het attributie-risico overschat (onderschat) wordt wanneer een positieve (negatieve) correlatie tussen de schadelijke factoren bestaat.

Een formule voor de geschatte variantie van AR bij grote steekproeven werd gegeven door Walter (1976). In vereenvoudigde notatie (vgl. Nijssse (1992)):

$$\hat{V}_{AR} = \frac{N_{..} N_{21} [N_{21} (N_{12} N_{21} - N_{11} N_{22}) + N_{..} N_{11} N_{22}]}{(N_{2.} N_{.1})^3} \quad (19)$$

Varianten op deze formule kan men vinden in Fleiss (1981, p. 76) en Leung en Kupper (1981). Voor de berekening van een betrouwbaarheidsinterval op basis van kleine steekproeven kan gebruik worden gemaakt van de procedure van Copas en Loeber (1990, pp. 304-305).

Voor het Schooloordeel is de puntschatting van AR .39. De onder- en bovengrens van het .95 -betrouwbaarheidsinterval zijn .30 resp. .48. Voor de GALO-test zijn de overeenkomstige resultaten .75 resp. .70 - .80.

In plaats van het attributie-risico wordt soms het relatieve risico (RR) gebruikt:

$$\text{RR} = \frac{p(B|A)}{p(B|\bar{A})} = \frac{\text{TPV}}{\text{ONV}} \quad (20)$$

Een nadeel van deze maat is dat voor een niet-waardeloze test de bovengrens ∞ is, zodat

een directe interpretatie moeilijk is. Uiteraard is AR in termen van RR te schrijven. Het verband tussen de aposteriori-kans $p(B|A)$ en het relatieve risico bij variërende basisverhoudingen, sensitiviteiten en specificiteiten werd bestudeerd door Murphy (1983). Zoals door hem werd aangetoond (en kan worden afgeleid uit (20)), kan een grote waarde van RR samengaan met een kleine waarde van TPV en omgekeerd. Verder wees Murphy erop dat de interpretaties van betrouwbaarheidsintervallen voor RR ook toepasbaar zijn op betrouwbaarheidsintervallen voor $\frac{Se}{(1 - Sp)}$. Wanneer dit laatste interval de waarde 1 bevat, betekent dit dat in de populatie $I = 0$ tot de mogelijkheden behoort.

5. Gebruik van (meer dan) twee tests

Wanneer een onderzoek(st)er overweegt naast een reeds in gebruik zijnde test één of meer nieuwe tests te gebruiken, rijst de vraag of afname van de nieuwe test(s) voldoende informatie oplevert om gebruik ervan te rechtvaardigen. Bij continu- en normaal-verdeelde predictoren en criteria wordt het antwoord op de vraag verkregen door het berekenen van de toename in R^2 , de gekwadrateerde meervoudige correlatiecoëfficiënt. Over de relatie tussen R^2 en de samenstelling van een optimale testbatterij bestaat zeer veel literatuur.

In het kader van Bayesiaanse kansrekening heeft de gestelde vraag minder aandacht gekregen. In de psychologische literatuur gaf Winkler (1982) als één van de weinigen een voorbeeld van de toename van $p(B|A)$ bij het afnemen van een tweede test. Uit Winkler's procedure is niet direct duidelijk dat de apriori-kans (basisverhouding) van de tweede test gevormd wordt door de aposteriori-kans van de eerste test.

In de medische literatuur wordt vaker aan het gebruik van twee tests gerefereerd. Een uitvoerig voorbeeld van veranderingen in de waarden van diagnostische maten als gevolg van afname van een tweede test werd gegeven door Weinstein en Fineberg (1980, pp. 125-136). Voor een meer systematische behandeling van dit thema zie men Jones en McClatchey (1988). Zoals op basis van ervaringen met de toename van R^2 verwacht mag worden, stijgt $p(B|A)$ minder sterk wanneer de diagnoses positief gecorreleerd zijn dan wanneer zij onafhankelijk zijn. In tegenstelling tot wat men misschien zou verwachten, kunnen de sensitiviteit, specificiteit en aposteriori-kansen soms dalen als gevolg van toevoeging van een tweede test.

Voor de gegevens over schoolkeuze wordt eerst uitgegaan van het Schooloordeel als

belangrijkste diagnostisch gegeven. Als men wil weten welke invloed toevoeging van de GALO-test op de diagnostische maten heeft, moet men twee nieuwe 2 x 2 -tabellen maken. De eerste tabel heeft betrekking op alle leerlingen met een positief Schooloordeel; de basisverhouding is de terecht-positieven-verhouding van het Schooloordeel. Met behulp van deze tabel worden de nieuwe sensitiviteit, terecht-positieven-verhouding en het attributie-risico berekend. De tweede tabel heeft betrekking op alle leerlingen met een negatief Schooloordeel; het complement van de basisverhouding is de terecht-negatieven-verhouding van het Schooloordeel. Deze tabel dient voor de berekening van de specificiteit en de terecht-negatieven-verhouding.

Toevoeging van de GALO-test blijkt op de meeste diagnostische maten een positieve invloed te hebben. De nieuwe selectieverhouding (.82) is beter in evenwicht met de nieuwe basisverhouding (.86). De nieuwe sensitiviteit is .90, een verbetering t.o.v. de oude waarde van .83. De specificiteit verandert niet en heeft ook nu de waarde .82. Het meest opvallend is de stijging van de terecht-positieven-verhouding van .54 naar .93. Vergeleken met de nieuwe basisverhouding is de stijging .07. De nieuwe terecht-negatieven-verhouding is .96, slechts .01 hoger dan de oude waarde. Ten opzichte van het complement van de basisverhouding is de stijging .08. Het nieuwe attributie-risico is .42, een forse daling t.o.v. .75 (GALO-test), maar een geringe verbetering t.o.v. .39 (Schooloordeel).

Bennett (1972) ontwikkelde toetsen voor de gelijkheid van sensitiviteit resp. specificiteit van (meer dan) twee diagnostische instrumenten, afgenomen aan dezelfde steekproef. In het geval van twee instrumenten dienen de gegevens te worden getabelleerd als in Tabel 1; voor de gebruikte symbolen zie men Tabel 7. De toetsen zijn toepassingen van McNemar's chi-kwadraat-toets voor de gelijkheid van twee afhankelijke proporties en hebben de vormen

$$\underline{X}^2 = \frac{(\underline{N}_{21} - \underline{N}_{31})^2}{(\underline{N}_{21} + \underline{N}_{31})} \quad \text{en} \quad \underline{X}^2 = \frac{(\underline{N}_{22} - \underline{N}_{32})^2}{(\underline{N}_{22} + \underline{N}_{32})} \quad (21)$$

voor sensitiviteit resp. specificiteit. De toetsingsgrootheden zijn asymptotisch chi-kwadraat-verdeeld met 1 vrijheidsgraad.

Toetsing van de sensitiviteiten van het Schooloordeel (.46) en de GALO-test (.83) geeft $\underline{X}^2 = 31.04$ ($p < .01$). Voor de specificiteiten (.98 resp. .82) is $\underline{X}^2 = 58.30$ ($p < .01$).

Tabel 7. Uitkomsten van een onderzoek met twee dichotome predictoren en een dichotoom criterium.

Diagnose		Uitkomst		
<u>A</u> ₁	<u>A</u> ₂	<u>B</u>	<u>B</u>	Tot.
+	+	<u>N</u> ₁₁	<u>N</u> ₁₂	<u>N</u> _{1.}
+	-	<u>N</u> ₂₁	<u>N</u> ₂₂	<u>N</u> _{2.}
-	+	<u>N</u> ₃₁	<u>N</u> ₃₂	<u>N</u> _{3.}
-	-	<u>N</u> ₄₁	<u>N</u> ₄₂	<u>N</u> _{4.}
		-----	-----	-----
Tot.		<u>N</u> _{.1}	<u>N</u> _{.2}	<u>N</u> _{..}

Door Bennett (1985) werden ook toetsen ontwikkeld voor de gelijkheid van aposteriorikansen voor (meer dan) twee tests, afgenomen aan dezelfde steekproef. In het geval van twee tests heeft de toets voor een positieve testuitslag de vorm:

$$\underline{X}^2 = \frac{[(N_{21} - N_{31})(N_{12} + N_{32}) - (N_{22} - N_{32})(N_{11} + N_{31})]^2}{[N_{11}(N_{22} - N_{32})^2 + N_{21}(N_{12} + N_{32})^2 + N_{31}(N_{12} + N_{22})^2] + [N_{22}(N_{11} + N_{21})^2 + N_{32}(N_{11} + N_{31})^2 + N_{42}(N_{21} - N_{31})^2]} \quad (22)$$

Voor een negatieve testuitslag gebruike men:

$$\underline{X}^2 = \frac{[(N_{21} - N_{31})(N_{22} + N_{42}) - (N_{22} - N_{32})(N_{21} + N_{41})]^2}{[N_{11}(N_{22} - N_{32})^2 + N_{21}(N_{22} + N_{42})^2 + N_{31}(N_{32} + N_{42})^2] + [N_{22}(N_{11} + N_{21})^2 + N_{32}(N_{21} + N_{41})^2 + N_{42}(N_{21} - N_{31})^2]} \quad (23)$$

Ook deze toetsingsgrootheden zijn asymptotisch chi-kwadraat-verdeeld met 1 vrijheidsgraad.

De terecht-positieven-verhoudingen van het Schooloordeel en de GALO-test zijn .86

resp. .54. Toepassing van (22) op de gegevens van Tabel 1 geeft $\underline{X}^2 = 6.51$ ($p < .02$). Voor de terecht-negatieven-verhoudingen (.88 resp. .95) wordt met (23) $\underline{X}^2 = 16.94$ gevonden.

Het is belangrijk op te merken dat (22) en (23) in verschillende vorm ook door Bennett (1972) werden gegeven. De aldaar vermelde $\underline{X}^2$ -waarden verschillen van de $\underline{X}^2$ -waarden in Bennett (1985), hoewel zij op hetzelfde voorbeeld betrekking hebben. De verschillende vormen zijn waarschijnlijk asymptotisch equivalent en de p -waarden zijn dan ook vergelijkbaar; vermoedelijk genieten de toetsen in Bennett (1985) in het algemeen de voorkeur (Snijders, 1991).

Soms komt het voor dat men een bestaande test wil vervangen door een nieuwe (betere) test. Een vergelijking van de (on)terecht-positieven- en (on)terecht-negatieven-verhoudingen van de oude en nieuwe test is dan belangrijk. Thibodeau (1981) en Vacek (1985) gaven methoden om deze vergelijking te maken.

6. Slotbeschouwing

Met de bespreking van enkele begrippen en technieken uit de medische statistiek is getracht aan te tonen dat een meer volledig en genuanceerd beeld van de diagnostische waarde van tests kan worden gegeven dan op basis van uitsluitend aposteriori-kansen mogelijk is. Zo is o.m. duidelijk gebleken dat de kans op een juiste positieve diagnose weliswaar groot kan zijn, maar dat dit resultaat gerelativeerd moet worden als de "verticale invalshoek" een lage sensitiviteit aan het licht brengt (voor Schooloordeel geldt een dergelijke uitkomst). Kennis van de sensitiviteit en specificiteit van een test zou eigenlijk aan het voorspellend gebruik vooraf moeten gaan, zoals uit de bespreking van de J -index is gebleken.

Een belangrijke ontwikkeling is de totstandkoming van inferentieel-statistische technieken, waardoor resultaten van meerdere onderzoeken met elkaar vergeleken kunnen worden en de keuze van te gebruiken tests beter kan worden gefundeerd. Wanneer men overweegt (meer dan) twee tests af te nemen, moet men bedacht zijn op kleine aantallen personen per combinatie van testuitslagen. Uit de gegevens van Tabel 1 blijkt bijvoorbeeld dat de keuze VWO in de helft van de gevallen goed voorspeld wordt voor de tien leerlingen waarvoor het Schooloordeel positief en de GALO-test negatief is. Een dergelijke

uitkomst behoeft kruisvalidatie. Merk verder op dat bij de combinatie Schooloordeel negatief - GALO-test positief de meeste leerlingen geen VWO-opleiding volgen. Bij de voorspellende waarde van de GALO-test kan een vraagteken worden gezet wanneer de uitslag niet overeenstemt met het Schooloordeel. Ook m.b.t. de winstpercentages .07 en .08 voor de terecht-positieven- resp. terecht-negatieven-verhouding, die toevoeging van de GALO-test aan het Schooloordeel opleveren, geldt dat kruisvalidatie gewenst is.

Tabel 8. Waarden diagnostische maten voor schoolkeuze VWO op basis van Schooloordeel (O), GALO-test (G), GALO-test gegeven Schooloordeel (G|O) en Schooloordeel gegeven GALO-test(O|G). Voor O en G: onder- en bovengrens .95-betrouwbaarheidsinterval onder vermelde waarde toegevoegd.

	<u>BV</u>	<u>SV</u>	<u>Se</u>	<u>Sp</u>	<u>TPV</u>	<u>TNV</u>	<u>AR</u>
<u>O</u>	.20	.11	.46	.98	.86	.88	.39
	.17	.08	.36	.97	.75	.84	.30
	.23	.13	.55	.99	.92	.90	.48
<u>G</u>	.20	.31	.83	.82	.54	.95	.75
	.17	.27	.76	.78	.46	.92	.70
	.23	.35	.90	.86	.61	.97	.80
<u>G</u> <u>O</u>	.86	.82	.90	.82	.93	.96	.42
<u>O</u> <u>G</u>	.54	.28	.49	.99	.93	.96	.29

X²-waarden toetsing sensitiviteiten, specificiteiten, terecht-positieven- en terecht-negatieven-verhoudingen: Se: $\chi^2 = 31.04$ ($p < .01$); Sp: $\chi^2 = 58.30$ ($p < .01$);

TPV: $\chi^2 = 6.51$ ($p < .02$); TNV: $\chi^2 = 16.94$ ($p < .01$).

In de vorige paragraaf is, aansluitend bij de gangbare praktijk, het Schooloordeel als belangrijkste gegeven beschouwd; aan de GALO-test werd een aanvullende rol toegekend. Men zou ook de GALO-test als eerste gegeven kunnen kiezen. Hoewel deze rolwisseling

geen verschil maakt voor de uiteindelijke aposteriori-waarden, bedenke men dat deze op basis van andere uitgangswaarden (basisverhouding enz.) worden verkregen. De meest spectaculaire verandering doet zich voor bij de terecht-positieven-verhouding: deze stijgt van .54 (GALO-test) naar .93 (GALO-test en Schoolordeel samen). Voor een gedetailleerde vergelijking van de resultaten die met beide volgorden worden verkregen, vergelijk men de regel G | O met regel O | G in Tabel 8, waarin alle uitkomsten zijn samengevat.

Literatuur

- Asch, D. A., Patton, J. D., & Hershey, J. C. (1990). Knowing for the sake of knowing: The value of prognostic information. Medical Decision Making, 10, 47-57.
- Asch, D. A., Patton, J. D., & Hershey, J. C. (1991). Prognostic information versus accuracy: Once more with meaning. Medical Decision Making, 11, 45-47.
- Baron, J. (1988). Thinking and deciding. Cambridge (U.K.): Cambridge University Press.
- Bennett, B. M. (1972). On comparisons of sensitivity, specificity and predictive value for a number of diagnostic procedures. Biometrics, 28, 793-800.
- Bennett, B. M. (1985). On tests for equality of predictive values for t diagnostic procedures. Statistics in Medicine, 4, 535-539.
- Bross, I. (1954). Misspecification in 2 x 2 tables. Biometrics, 10, 478-486.
- Connell, F. A., & Koepsell, T. D. (1985). Measures of gain in certainty from a diagnostic test. American Journal of Epidemiology, 121, 744-753.
- Copas, J. B., & Loeber, R. (1990). Relative improvement over chance (RIOC) for 2 x 2 tables. British Journal of Mathematical and Statistical Psychology, 43, 293-307.
- Diamond, G. A. (1991). Point of information. Medical Decision Making, 11, 42-44.
- Diamond, G. A., & Forrester, J. S. (1983). Metadiagnosis; An epistemological model for clinical judgment. The American Journal of Medicine, 75, 129-137.
- Evans, J. St. B. T. (1990). Bias in human reasoning; Causes and consequences. London: Erlbaum.
- Fleiss, J. L. (1981). Statistical methods for rates and proportions (2nd ed.). New York: Wiley.
- Heckering, P. S. (1988). Confidence in diagnostic testing. Journal of General Internal Medicine, 3, 604-606.

- Jones, R. H., & McClatchey, M. W. (1988). Beyond sensitivity, specificity and statistical independence. Statistics in Medicine, 7, 1289-1295.
- Kema, G. N., & Kouwer, B. J. (1958). G.A.L.O.; Groninger Afsluitingsonderzoek Lager Onderwijs. Intelligentietest zesde klas lagere school. Handleiding. Groningen: Groninger Instituut voor Toegepaste Psychologie en Psychotechniek.
- Leung, H. M., & Kupper, L. L. (1981). Comparisons of confidence intervals for attributable risk. Biometrics, 37, 293-302.
- Monsour, M. J., Evans, A. T., & Kupper, L. L. (1991). Confidence intervals for post-test probability. Statistics in Medicine, 10, 443-456.
- Murphy, J. R. (1983). The relationship of relative risk and positive predictive value in 2 x 2 tables. American Journal of Epidemiology, 117, 86-89.
- Nijse, M. (1992). Attributable risk and relative improvement over chance: A note on Copas and Loeber (1990). British Journal of Mathematical and Statistical Psychology, 45, 329-333.
- Nijse, M. (1994). De predictieve waarde van schooloordeel en intelligentietest voor schoolkeuze en schoolsucces in het voortgezet onderwijs. (In voorbereiding).
- Quade, D., Lachenbruch, P. A., Whaley, F. S., McClinch, B. K., & Haley, R. W. (1981). Effects of misclassifications on statistical inferences in epidemiology. American Journal of Epidemiology, 111, 503-515.
- Snijders, T. A. B. (1991). Persoonlijke mededeling.
- Thibodeau, L. A. (1981). Evaluating diagnostic tests. Biometrics, 37, 801-804.
- Vacek, P. M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. Biometrics, 41, 959-968.
- Vecchio, T. J. (1966). Predictive value of a single diagnostic test in unselected populations. New England Journal of Medicine, 274, 1171-1174.
- Walter, S. D. (1976). Estimation and interpretation of attributable risk in health research. Biometrics, 32, 829-849.
- Walter, S. D. (1978). Calculation of attributable risk from epidemiological data. International Journal of Epidemiology, 7, 175-182.
- Weinstein, M. C., & Fineberg, H. V. (1980). Clinical decision analysis. Philadelphia: Saunders.
- Winkler, R. L. (1982). The Bayesian approach: A general review. In G. Keren (Ed.), Statistical and methodological issues in psychological and social science research

(pp. 217-244). Hillsdale, NJ: Erlbaum.

Yerushalmy, J. (1947). Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray technique. Public Health Reports, 62, 1432-1449.

Youden, W. J. (1950). Index for rating diagnostic tests. Cancer, 3, 32-35.

Ontvangen: 18-5-1992

Geaccepteerd: 3-12-1993