VIDENSKAB OG PRAKSIS | OVERSIGTSARTIKEL

Mål for diagnostiske tests ydeevne

Klinisk assistent Bo Rud, overlæge Peter Matzen & lektor Jørgen Hilden

H:S Hvidovre Hospital, Endokrinologisk Afdeling, Osteoporoseenheden og Gastroenterologisk Afdeling, og Københavns Universitet, Institut for Folkesundhedsvidenskab, Biostatisk Afdeling

Resume

Evidensgrundlaget for flere diagnostiske test er ikke fyldestgørende, og det har vist sig, at klinikere kan have svært ved at fortolke mål for diagnostiske tests ydeevne. Ud fra et konkret eksempel beskrives gængse mål for en tests ydevne (sensitivitet, specificitet, *likelihood*-ratioer samt positiv og negativ prædiktiv værdi). En tests ydeevne angives ofte som sensitivitet og specificitet, men disse mål er først relevante for klinikeren, når de omregnes til prædiktive sandsynligheder. Det vises, hvordan man bestemmer disse sandsynligheder, dels ved beregning, dels ved brug af *likelihood*-ratioer og Fagans nomogram. Fortolkningen af resultater fra undersøgelser af diagnostiske tests ydeevne belyses med udgangspunkt i: 1) deltagernes placering i det kliniske spektrum, 2) undersøgelsens metodologiske kvalitet samt 3) hvor sikkert målene for ydeevnen er bestemt. Det beskrives, hvordan det kliniske spektrum og metodologisk kvalitet kan påvirke en tests træfsikkerhed, og metoder til estimering af usikkerhed anvises.

Som led i bestræbelserne på at gøre det kliniske arbejde evidensbaseret har der igennem de senere år været tiltagende fokus på undersøgelser af diagnostiske tests træfsikkerhed. Dette hænger sammen med, at man har konstateret, at evidensgrundlaget for mange diagnostiske test ikke er fyldestgørende [1, 2]. For at rette op på denne situation er der udarbejdet retningslinjer for rapportering og design af undersøgelser af diagnostiske test i form af Standards for Reporting of Diagnostic Accuracy (STARD) [3], og systematiske oversigtsartikler over diagnostiske test ydeevne vil fremover indgå som en del af Cochrane-samarbejdet [4]. Den øgede interesse for evidensgrundlaget for diagnostiske test er nødvendig, da udredning vha. træfsikre diagnostiske test er en forudsætning for relevant behandling og pålidelig prognostik.

Der er imidlertid også andre problemer forbundet med diagnostiske test, idet det har vist sig, at klinikere kan have svært ved at fortolke og anvende den information, der beskriver ydeevnen en test [5, 6].

I denne oversigtsartikel vil vi først gennemgå gængse, klinisk relevante mål for diagnostiske tests ydeevne. Vi vil derefter præsentere læseren for en række forhold vedr. design og rapportering, som det er nødvendigt at have kendskab til, når man vurderer forskningsresultater, der omhandler ydeevnen af diagnostiske test. Diagnostisk test skal i denne artiklen forstås i bred forstand, dvs. som test til screening, diagnostik og monitorering.

Sensitivitet, specificitet, positiv og negativ prædiktiv værdi

Vi vil beskrive målene for diagnostisk træfsikkerhed ved hjælp af et klinisk relevant eksempel. Vi har valgt at benytte Ottawa Ankle Rule (OAR) [7, 8], der er en simpel screeningstest udviklet mhp. at udelukke immobiliseringskrævende skade (malleollær fraktur eller avulsion >3 mm) efter et ankeltraume (**Figur 1**). Referencetesten, som OAR skal vurderes i forhold til, er et røntgenbillede af anklen.

Da *Auleley et al* [9] validerede OAR på 357 konsekutivt udvalgte patienter, der henvendte sig med et ankeltraume til en skadestue i Paris, opnåede de resultaterne i **Tabel 1**. Man benytter vanligvis fire mål til at beskrive en tests træfsikkerhed (*accuracy*): Sensitivitet, specificitet, positiv og negativ prædiktiv værdi [10].

Sensitiviteten angiver andelen af syge patienter, der har en positiv test, her andelen af patienter med røntgenverificeret fraktur, som opfylder OAR:

$$\text{Sensitivitet} = SP/(SP + FN) = 48/49 = 0{,}98.$$

Specificiteten betegner andelen af raske, der har en negativ test, her andelen af patienter uden fraktur, som ikke opfylder OAR:

$$\text{Specificitet} = SN/(SN + FP) = 137/308 = 0{,}44.$$

For klinikeren er det imidlertid ikke så relevant at kende sandsynligheden for, at OAR er opfyldt (eller ikke opfyldt), når patienten har en fraktur (eller ikke har en fraktur). Derimod er det relevant at have kendskab til, hvordan sandsynligheden for fraktur eller ikkefraktur afhænger af udfaldet af OAR.

Blandt dem, der opfylder OAR, kaldes den andel, som faktisk har en fraktur, for testens positive prædiktive værdi (PPV):

$$PPV = SP/(SP + FP) = 48/219 = 0{,}22.$$

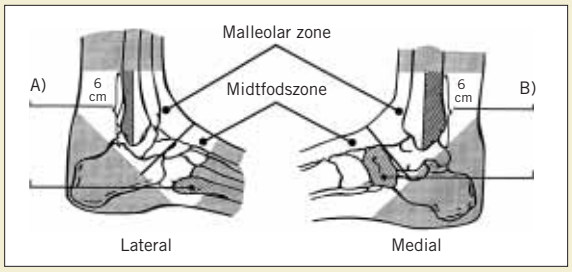
Blandt dem, der ikke opfylder OAR, kaldes den andel, som ikke har en fraktur, for testens negative prædiktive værdi (NPV):

$$NPV = SN/(SN + FN) = 137/138 = 0{,}99.$$

Sensitivitet og specificitet angiver med andre ord sandsynligheden for de testudfald, klinikeren kan forvente at finde, når

En røntgenundersøgelse af ankelregionen efter et ankeltrauma er nødvendig, hvis der er:

1) Smerter i anklen *og enten*
2) ømhed langs den posteriore kant eller ud for spidsen af malleolus lateralis eller medialis (se A og B nedenfor)
   eller
3) patienten er ude af stand til at støtte på foden umiddelbart efter traumet og ude af stand til at gå fire skridt i skadestuen



Figur 1. Ottawa Ankle Rule. Gengivet med tilladelse fra *Ian Stiell* (for yderligere information om Ottawa Ankle Rule, se http://www.ohri.ca/programs/clinical_epidemiology/OHDEC/ankle_rule/flash_ankle_rule.htm).

Tabel 1. Fransk valideringsstudie af Ottawa Ankle Rule (OAR) blandt 357 patienter med et ankeltraume.

	Fraktur	Ikke fraktur	Total
OAR opfyldt	48 (SP)	171 (FP)	219
OAR ej opfyldt	1 (FN)	137 (SN)	138
Total	49	308	357

SP: sandt positive; FP: falsk positive; FN: falsk negative; SN: sandt negative.

en given diagnose er kendt, hhv. afkræftet. I forbindelse med udredning er diagnosen imidlertid ikke kendt – kun testudfaldet er kendt. Derfor er det de prædiktive sandsynligheder (PPV, NPV), der er relevante for klinikeren. De angiver sandsynligheden for, at tilstanden, man tester for, er til stede (eller ikke til stede), når testudfaldet er kendt.

Det ses, at OAR's negative udfald er bedre til afkræftelse af fraktur, end det positive udfald er til bekræftelse af fraktur. I *Auleleys et al's* undersøgelse var hyppigheden af fraktur 49/357 = 0,14. Ved et positivt udfald øges sandsynligheden for fraktur fra 0,14 til 0,22, hvilket er en beskeden diagnostisk gevinst. Omvendt mindskes sandsynligheden for fraktur fra 0,14 til 0,007 (1–NPV = 1/138) ved et negativt testudfald, hvilket er en relevant diagnostisk gevinst, idet fraktur nu med rimelighed kan udelukkes. Da 138 patienter havde et negativt OAR-udfald, kunne OAR have reduceret antallet af røntgenbilleder med 39%, hvis man i øvrigt mener, at en sandsynlighed for fraktur på 1/138 er ubetydelig.

Alle de foregående vurderinger er naturligvis behæftet med sædvanlig statistisk stikprøveusikkerhed som drøftet nedenfor i afsnittet om usikkerhed.

Forventer man en anden prævalens end 0,14 i sin lokale population, men uændret sensitivitet og specificitet, kan man bestemme de prædiktive værdier vha. den beregning, der er gennemgået i ti trin i **Tabel 2**. Beregningerne illustrerer en vigtig pointe, nemlig at de prædiktive værdier afhænger af prævalensen af den tilstand, man tester for. Helt tilsvarende beregninger kan benyttes til bestemmelse af de prædiktive sandsynligheder i artikler, hvor kun sensitiviteten, specificiteten og prævalensen af tilstanden, man tester for, er angivet. Det skal understreges, at beregningerne forudsætter, at sensitiviteten og specificiteten kan ekstrapoleres frit fra *Auleleys et al*, hvilket ikke nødvendigvis er tilfældet. En metaanalyse over 27 valideringsstudier af OAR har imidlertid vist, at antagelsen holder for sensitiviteten, når OAR anvendes i stikprøver af skadestuepatienter med ankeltraumer. Den poolede sensitivitet var 97,6%, 95% konfidensinterval (KI) (96,4-98,9). Mht. specificiteten findes intet poolet estimat i metaanalysen, derimod var interkvartilspændvidden 0,25-0,45 [11].

Det lave antal falsk negative testudfald giver anledning til OAR's høje sensitivitet og NPV, hvilket som nævnt gør OAR velegnet til afkræftelse af fraktur. På tilsvarende vis kan en test med et lavt antal falsk positive udfald have høj specificitet og PPV, og af den grund kan testen være velegnet til at bekræfte tilstanden, man tester for. Desværre er det ikke altid sådan, at test med høj sensitivitet er velegnede til at afkræfte, mens test med høj specificitet er velegnede til at bekræfte [12]. Dette hænger sammen med de prædiktive værdiers prævalensafhængighed. Hvis sygdommen, man tester for, er sjælden (f.eks. cancer), kan man opnå høj NPV, selv om testens sensitivitet kun er moderat. Hvis omvendt tilstanden, man tester for, er meget hyppig, er et negativt testudfald upålideligt, selv når sensitiviteten er høj. Er tilstanden, man tester for, meget sjælden, så er et positivt testudfald stadig upålideligt, selv om specificiteten er i top (f.eks. hiv-test blandt patienter uden risikoadfærd).

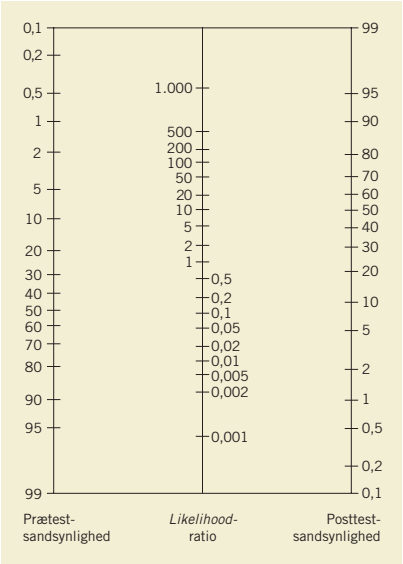
Likelihood-ratioer

En alternativ beskrivelse af OAR's ydeevne, som kan være let-

Tabel 2. Omregning i ti trin fra sensitivitet og specificitet til positiv og negativ prædiktiv værdi af Ottawa Ankle Rule (OAR) i et hypotetisk klientel bestående af 1.000 patienter, hvor 25% har fraktur, sensitiviteten = 0,98 og specificiteten = 0,44.

	Fraktur	Ikkefraktur	Total	Prædiktive værdier
OAR opfyldt	[3] 0,98 × 250 = 245	[6] 750-330 = 420	[7] 245 + 420 = 665	[9] PPV = 245/665 = 0,37
OAR ej opfyldt	[4] 250-245 = 5	[5] 0,44 × 750 = 330	[8] 5 + 330 = 335	[10] NPV = 330/335 = 0,99
Total	[1] 0,25 × 1.000 = 250	[2] 1.000-250 = 750	1.000	–

1-10: Angiver rækkefølgen af de ti trin i beregningen af prædiktive sandsynligheder ud fra sensitivitet, specificitet og sygdomsprævalens.



Figur 2. Nomogram til udregning af eftertest-sandsynlighed ud fra tilstandens prætestsandsynlighed (prævalens) og den til testudfaldet hørende *likelihood*-ratio (Copyright © 1975 Massachusetts Medical Society).

tere at håndtere i klinikken, består i at angive, hvor hyppigt et givet testudfald forekommer hos patienter med fraktur i forhold til hyppigheden af udfaldet hos patienter uden fraktur. Dette forhold benævnes testudfaldets *likelihood*-ratio [13].

Likelihood-ratioen ved det positive udfald af OAR (LR+) beregnes som:

$$LR+ = \text{sensitivitet}/(1-\text{specificitet}) = 0{,}98/(1-0{,}44) = 1{,}75.$$

I *Auleley et al's* stikprøve forekom et positivt OAR-udfald altså 1,75 gange hyppigere hos patienter med en ankelfraktur end hos patienter uden.

Tilsvarende angiver *likelihood*-ratioen for det negative udfald af OAR (LR–), hvor hyppigt et negativt udfald forekommer hos patienter med fraktur i forhold til hyppigheden af et negativt udfald hos patienter uden fraktur:

$$LR- = (1-\text{sensitivitet})/\text{specificitet} = (1-0{,}98)/0{,}44 = 0{,}045.$$

Et negativt testudfald ses altså 22 (1/0,045) gange hyppigere blandt patienter uden fraktur end blandt patienter med fraktur. LR+ og LR– bekræfter det, vi allerede vidste: Et negativt OAR-udfald afkræfter i høj grad fraktur, mens et positivt udfald ikke gør klinikeren meget klogere. Man kan sige, at *likelihood*-ratioer afspejler, i hvilket omfang et testudfald ændrer før-test-sandsynligheden for tilstanden, man tester for. En test med høj ydeevne har ekstreme LR-værdier, dvs. værdier som enten er væsentligt lavere end 1,0 (<0,1-0,2) eller væsentligt højere (>5-10).

Likelihood-ratioer kan benyttes til beregning af prædiktive sandsynligheder, dvs. sandsynligheden for, at patienten har tilstanden, man tester for, når testudfaldet er positivt (PPV) hhv. negativt (1-NPV). Beregningerne tager udgangspunkt i den forventede hyppighed af tilstanden, man tester for, som skal udtrykkes som en odds. I *Auleleys et al's* stikprøve var hyppigheden af ankelfraktur 14%, hvilket omregnes til odds:

$$\text{odds}_{\text{før test}} = \text{prævalens}/(1-\text{prævalens}) = 0{,}14/(1-0{,}14) = 0{,}16.$$

Ganges $\text{odds}_{\text{før test}}$ med LR+ opnås odds for fraktur efter test:

$$\text{odds}_{\text{efter test}} = (LR+) \times (\text{odds}_{\text{før test}}) = 1{,}75 \times 0{,}16 = 0{,}28.$$

$\text{Odds}_{\text{efter test}}$ kan herefter omregnes til PPV:

$$PPV = \text{odds}_{\text{efter test}}/(\text{odds}_{\text{efter test}} + 1) = 0{,}28/(0{,}28 + 1) = 0{,}22.$$

Den tilsvarende beregning med LR– giver 1-NPV. Beregningerne er imidlertid tidskrævende, men en væsentlig pointe ved *likelihood*-ratioer er, at man vha. nomogrammet i **Figur 2** helt kan slippe for at regne [14]. Lægges en ret linje gennem 14%-punktet på før-test-aksen til venstre og videre gennem 1,75-punktet på *likelihood*-aksen i midten, skærer linjen efter-test-aksen til højre lige over 20%-punktet og det bekræftes, at sandsynligheden for fraktur, når OAR er positiv, er 22%. Trækker man en linje gennem 14%'s punktet på aksen til venstre og videre gennem 0,045 (LR–)-punktet på *likelihood*-aksen i midten, skærer linjen efter-test-aksen til højre nær 1%-punktet. Sandsynligheden for fraktur, når OAR er negativ er altså 0,01, hvilket netop er 1-NPV.

Likelihood-ratioer er især nyttige, når en test har flere end to mulige udfald. Testudfaldet for kvantitative test og laboratorieprøver ranginddeles sædvanligvis i en række kategorier, men også kvalitative test har ofte flere end to mulige udfald, f.eks. inden for billeddiagnostik, hvor testudfaldet kan graderes som sikkert abnormt, muligt abnormt eller normalt. Her udregnes en *likelihood*-ratio for hvert af udfaldene, LR(udfald) (**Tabel 3**).

Med nomogrammet ved hånden samt kendskab til LR(udfald) og prævalensen lettes beregningen af de klinisk relevante prædiktive sandsynligheder (PPV, NPV) betydeligt. Den diagnostiske gevinst ved brug af testen fremstår klart, når nomogrammet anvendes. Sandsynligheden for sygdom før testen er angivet på før-test-aksen til venstre, mens efter-test-sandsynligheden er angivet på aksen til højre.

Tabel 3. Beregning af *likelihood*-ratioer, for en test med fire mulige udfald W, X, Y, Z.

	Referencetest		
Testresultat	**abnorm**	**normal**	**Likelihood-ratio**
W	a	b	LR(W) = (a/n)/(b/m)
X	c	d	LR(X) = (c/n)/(d/m)
Y	e	f	LR(Y) = (e/n)/(f/m)
Z	g	h	LR(Z) = (g/n)/(h/m)
Total	n = a+c+e+g	m = b+d+f+h	–

Vurdering af undersøgelser af diagnostiske test ydeevne

Når klinikeren fortolker studier af diagnostiske test ydeevne, er der tre væsentlige forhold, der bør tages i betragtning, før det vurderes, om resultaterne er klinisk relevante. Dels må klinikeren vurdere, om patienterne i undersøgelsen svarer til dem, hun ser til daglig, dels må hun vurdere, om resultaterne er valide, og endelig må hun vurdere usikkerheden på estimaterne for ydeevnen.

Spektrum

Det er velkendt, at træfsikkerheden af en diagnostisk test kan variere mellem undergrupper af patienter, afhængigt af hvor patienterne befinder sig i det kliniske spektrum [15]. Spektrum betegner her variationsbredden i de anamnestiske, kliniske og patologiske karakteristika, som patienter med en given sygdom har. Disse karakteristika varierer, bl.a. afhængigt af hvor fremskreden sygdommen er. For eksempel har sensitiviteten af arbejds-elektrokardiogram (EKG) sammenholdt med >75% stenose ved koronar arteriografi vist sig at stige med antallet af stenoserede koronarkar i nogle undersøgelser [16, 17]. Men også andre forhold kan påvirke målene for træfsikkerhed f.eks. køn, alder, etnicitet og sygdomsprævalens [15]. I nogle studier har man for eksempel fundet, at sensitiviteten af arbejds-EKG er højere for mænd end for kvinder og højere for ældre end for yngre [16, 17]. Det modsatte var tilfældet for specificiteten. Arbejds-EKG'ets ydeevne udtrykt ved *likelihood*-ratioer har også vist sig at variere mellem visse undergrupper af patienter [17].

På denne baggrund står det klart, at en tests ydeevne vil afhænge af sammensætningen af den stikprøve, som testen er afprøvet på. Er undergrupper, hvor testen f.eks. har lavere specificitet og højere sensitivitet overrepræsenteret i stikprøven, vil undersøgelsen give et skævt billede af testens ydeevne, hvis testen anvendes blandt patienter, hvor de pågældende undergrupper udgør en mindre andel. Denne skævhed betegnes spektrumbias. For eksempel kan estimaterne for træfsikkerheden af en test udført blandt indlagte patienter med fremskreden sygdom ikke antages at være gældende for patienter i almen praksis, hvor patienter hovedsagligt vil have den pågældende sygdom i et tidligt stadium, og hvor banale symptomårsager er hyppige [18].

For at klinikeren kan regne med, at estimater for ydeevnen af en test er overførbare fra en undersøgelse til klinisk praksis, må hun altså stå med en gruppe af patienter, som mht. placeringen i det kliniske spektrum svarer til patienterne i de stikprøver, testen er evalueret på.

Validitet

For undersøgelser af ydeevnen af diagnostiske test gælder en række metodologiske krav, som skal være opfyldt, før undersøgelsens konklusioner kan tages for pålydende [19]. Disse krav er opsummeret i STARD-initiativets checkliste [3]. De væsentligste metodekrav er gengivet i stikordsform i **Figur 3**.

- Den kliniske problemstilling, testen skal afklare, er velbeskrevet
- Patientmaterialet, rekrutteringen og det kliniske miljø er velbeskrevet og relevant for problemstillingen
- Der er overvejelser vedr. stikprøvestørrelse
- Test og referenceundersøgelse er velbeskrevne mht. tekniske aspekter, udførelse og kategorisering af udfald
- Det er angivet, hvem og hvor mange der udførte test, og referenceundersøgelse med oplysninger om uddannelse og kompetence
- Referenceundersøgelsen er foretaget på alle patienter uanset udfaldet af testen
- De(n), der vurderer udfaldet af testen, er blindet mht. udfaldet af referenceundersøgelsen
- De(n), der vurderer udfaldet af referenceundersøgelsen, er blindet mht. udfaldet af testen
- Intra- og interobservatørvariation er opgjort for såvel test som referenceundersøgelse

Figur 3. Metodologiske krav til undersøgelser af diagnostiske tests ydeevne.

I en metaanalyse er det dokumenteret, at målene for diagnostisk træfsikkerhed overestimeres i undersøgelser med metodologiske mangler [20].

Usikkerhed

Estimater for træfsikkerhed og *likelihood*-ratioer bør angives med f.eks. 95% sikkerhedsintervaller, så klinikeren er i stand til at vurdere usikkerheden på estimaterne. I en undersøgelse af tredje hjertelyds træfsikkerhed mht. hjerteinsufficiens konstateret ved ekkokardiografi (*fractional shortening* <25%), hvor 16% af deltagerne havde hjerteinsufficiens, fandt man eksempelvis, at PPV var 0,77 (95% KI: 0,46-0,95) og NPV var 0,87 (95% KI: 0,83-0,91) [21].

Sikkerhedsintervallet for PPV er meget bredt og gør det vanskeligt at vurdere, om testen er klinisk relevant.

Er sikkerhedsintervaller ikke angivet, kan de relativt let beregnes (**Figur 4**) [13].

Sammenfatning

Vi har beskrevet en række mål for ydeevnen af diagnostiske test. Det fremgår, at sensitivitet og specificitet ikke er umiddelbart relevante for klinikeren, medmindre de omregnes til prædiktive sandsynligheder (PPV, NPV) eller *likelihood*-ratioer. I omregningen til prædiktive sandsynligheder skal man huske at tage højde for den lokale prævalens af tilstanden, man tester for, hvis denne er kendt. Beregninger kan undgås ved brug af *likelihood*-ratioer og Fagans nomogam. Uanset hvilke mål for ydeevnen, man finder i tidsskriftsartikler, bør man nøje overveje deres overførbarhed, validitet og den statistiske usikkerhed, der er knyttet til dem, før målene lægges til grund for kliniske beslutninger. Når man har vurderet, at resultaterne vedr. en ny diagnostisk test er valide og overførbare til en given klinisk sammenhæng, og at usikkerheden er acceptabel, står det tilbage at vurdere, om resultaterne er klinisk relevante. I tilfælde, hvor testen udgør et nyt tilbud til patien-

For 2 × 2-tabellen:

	Reference (+)	Reference (–)	Total
Test (+)	*a*	*b*	*a+b*
Test (–)	*c*	*d*	*c+d*
Total	*a+c*	*b+d*	*a+b+c+d*

95% sikkerhedsgrænser på estimatet for træfsikkerhed, *p* (sensitivitet, specificitet, positiv og negativ prædiktiv værdi)

$p \pm 1{,}96 \times \mathrm{SE}(p)$, hvor $\mathrm{SE}(p) = \sqrt{\frac{p(1-p)}{n}}$

For sensitiviteten er *p* eksempelvis: $a/(a+b)$, hvor $n = a+b$

Denne tilnærmede formel forudsætter: $np > 10$ samt $n(1-p) > 10$

95% sikkerhedsgrænser for estimatet for *likelihood*-ratioen for en positiv hhv. en negativ test:

$$\mathrm{LR}(+) \times \exp\left[\pm 1{,}96\sqrt{\left(\frac{1}{a}-\frac{1}{a+c}\right)+\left(\frac{1}{b}-\frac{1}{b+d}\right)}\right]$$

$$\mathrm{LR}(-) \times \exp\left[\pm 1{,}96\sqrt{\left(\frac{1}{c}-\frac{1}{a+c}\right)+\left(\frac{1}{d}-\frac{1}{b+d}\right)}\right]$$

For 2 × k-tabellen:

Testudfald	Reference (+)	Reference (–)	*Likelihood*-ratio
W	*a*	*b*	$\mathrm{LR(W)} = (a/n_+)/(b/n_-)$
X	*c*	*d*	$\mathrm{LR(X)} = (c/n_+)/(d/n_-)$
Y	*e*	*f*	$\mathrm{LR(Y)} = (e/n_+)/(f/n_-)$
Z	*g*	*h*	$\mathrm{LR(Z)} = (g/n_+)/(h/n_-)$
Total	n_+	n_-	–

95% sikkerhedsgrænser for estimatet for *likelihood*-ratioen for testudfaldene W, X, Y, Z:

$$\exp\left[\ln\frac{p_+}{p_-} \pm 1{,}96\sqrt{\frac{1-p_-}{n_+p_+}+\frac{1-p_-}{n_-p_-}}\right]$$

Eksempel: For testudfaldet W: $p_+ = a/n_+$ og $p_- = b/n_-$

Figur 4. Beregning af 95% sikkerhedsintervaller på estimater for træfsikkerhed og *likelihood*-ratioer.

terne, afhænger denne vurdering af en værdidom, der dels involverer ydeevnen, men også etiske og økonomiske aspekter. I tilfælde, hvor den nye test skal erstatte eller supplere en eksisterende test, må beslutningen hvile på undersøgelser, hvori man sammenligner ydeevnen af den nye og den gamle test, eller sammenligner kombinationen af den nye og den gamle test med den gamle test alene. Herudover må man også inddrage forhold ved testen, såsom ubehag for patienten, pris og hvor let, den kan udføres. I alle tilfælde må det som minimum kræves, at diagnostiske test giver information, der påvirker behandlingsbeslutningen, så man opnår de bedst mulige konsekvenser for patienten [22].

Korrespondance: *Bo Rud*, Osteoporoseenheden 545, Endokrinologisk Afdeling, H:S Hvidovre Hospital, DK-2650 Hvidovre. E-mail: borud@mail.dk

Antaget: 22. september 2004
Interessekonflikter: Ingen angivet

Litteratur

1. Oosterhuis WP, Niessen RW, Bossuyt PM. The science of systematic reviewing studies of diagnostic tests. Clin Chem Lab Med 2000;38:577-88.
2. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. JAMA 1995;274:645-51.
3. Bossuyt PM, Reitsma JB, Bruns DE et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ 2003;326:41-4.
4. Deeks J, Gatsonis C, Bossuyt P et al. Cochrane reviews on diagnostic test accuracy – a progress report. Cochrane News 2004;31:1, 5.
5. Hoffrage U, Lindsey S, Hertwig R et al. Medicine. Communicating statistical information. Science 2000;290:2261-2.
6. Gigerenzer G. Reckoning with risk. Learning to live with uncertainty. 1 ed. London: Penguin Books, 2003.
7. Stiell IG, Greenberg GH, McKnight RD et al. A study to develop clinical decision rules for the use of radiography in acute ankle injuries. Ann Emerg Med 1992;21:384-90.
8. Stiell IG, Greenberg GH, McKnight RD et al. Decision rules for the use of radiography in acute ankle injuries. JAMA 1993;269:1127-32.
9. Auleley GR, Kerboull L, Durieux P et al. Validation of the Ottawa ankle rules in France: a study in the surgical emergency department of a teaching hospital. Ann Emerg Med 1998;32:14-8.
10. Sackett DL, Haynes RB, Guyatt GH et al. The interpretation of diagnostic data. Clinical Epidemiology. A basic science for clinical medicine. Boston: Little, Brown and Company, 1991:69-152.
11. Bachmann LM, Kolb E, Koller MT et al. Accuracy of Ottawa ankle rules to exclude fractures of the ankle and mid-foot: systematic review. BMJ 2003; 326:417.
12. Pewsner D, Battaglia M, Minder C et al. Ruling a diagnosis in or out with »SpPIn« and »SnNOut«: a note of caution. BMJ 2004;329:209-13.
13. Habbema JD, Eijkemans R, Krijnen P et al. Analysis of data on the accuracy of diagnostic tests. I: Knottnerus JA, ed. The evidence base of clinical diagnosis. London: BMJ Books, 2002:117-43.
14. Fagan TJ. Letter: Nomogram for Bayes theorem. N Engl J Med 1975;293: 257.
15. Whiting P, Rutjes AW, Reitsma JB et al. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004; 140:189-202.
16. Hlatky MA, Pryor DB, Harrell FE et al. Factors affecting sensitivity and specificity of exercise electrocardiography. Am J Med 1984;77:64-71.
17. Moons KG, van Es GA, Deckers JW et al. Limitations of sensitivity, specificity, likelihood ratio, and bayes' theorem in assessing diagnostic probabilities: a clinical example. Epidemiology 1997;8:12-7.
18. Van den Hoogen HM, Koes BW, van Eijk JT et al. On the accuracy of history, physical examination, and erythrocyte sedimentation rate in diagnosing low back pain in general practice. Spine 1995;20:318-27.
19. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. J Clin Epidemiol 2003;56:1118-28.
20. Lijmer JG, Mol BW, Heisterkamp S et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061-6.
21. Davie AP, Francis CM, Caruana L et al. Assessing diagnosis in heart failure: which features are any use? QJM 1997;90:335-9.
22. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. JAMA 1994;271:703-7.