3.1 **Utvärdering av diagnostiska metoder**

Sensitivitet och specificitet

Ett diagnostiskt tests prestanda anges med testets sensitivitet och specificitet. Uträkning av sensitivitet och specificitet förutsätter att man jämfört det nya testets utfall med någon standard eller referensmetod. Denna standard kan vara en mycket fullständig undersökning såsom t ex obduktion. I andra fall jämför man med ett annat test vars testegenskaper man känner och/eller som av tradition varit det sätt som en diagnos har etablerats på. I vissa situationer får man lita till en klinisk uppföljning av patienterna, t ex klinisk uppföljning av patienter som har ett ultraljudsfynd som inte talar för djup ventrombos för att se om kliniska symtom på trombos uppträder. Vi använder i rapporten termen managementstudie om undersökningar som på detta sätt verifierat att diagnosen inte föreligger.

I Faktaruta 1 visas utfallet från en studie som jämförde ”real time b-mode ultrasonography” med flebografi [5]. Resultatet för 220 patienter summerades i en tabell.

Sensitiviteten anger den andel av de sjuka där testet är positivt, och specificiteten den andel av de friska som testet kan frikänna. Detta är helt enkelt definitioner, men som får flera viktiga konsekvenser:

- Sensitiviteten beräknas endast bland de med sjukdom. Specificiteten å andra sidan beräknas enbart på de friska. Detta gör att sensitivitet och specificitet är oberoende av prevalensen av sjukdomen i den undersökta befolkningen. Emellertid kan olika egenskaper hos de sjuka påverka sensitiviteten, respektive hos de friska påverka specificiteten. Om man t ex undersöker en population symtomatiska patienter som huvudsakligen har kliniskt manifesta, utbredda tromboser, så kan

det hända att den beräknade sensitiviteten inte är överförbar till en situation där man screenar asymtomatiska postoperativa patienter som har tidiga eller små tromboser. Specificiteten för testet kan variera beroende på om de utan trombos (de friska) t ex är vanliga patienter från en akutmottagning eller om de är personer som i stor utsträckning tidigare haft tromboser.

- Sensitivitet och specificitet är mått som fås fram när vi redan har jämfört utfallet av det nya testet med facit från en referensmetod: Sensitivitet beskriver sannolikheten att testet är positivt givet att sjukdomen är närvarande. Specificitet beskriver sannolikheten att testet är negativt givet att sjukdomen inte är närvarande. I den kliniska vardagen är man emellertid oftast intresserad av ett annat perspektiv, nämligen frågan om hur stor sannolikheten är att sjukdomen är närvarande givet att testet är positivt. Motsvarande fråga för ett negativt test är: Vad är sannolikheten att patienten är fri från sjukdom givet att testet är negativt. Det förstnämnda kallas det positiva prediktiva värdet och det sistnämnda det negativa prediktiva värdet, definierade i Faktaruta 1.

- Eftersom sensitivitet och specificitet således beräknas i två olika populationer så adderas inte heller totalsumman upp till 1,0 (eller till 100 procent). Om vi studerar ett test där gränsen för vad som anses patologiskt respektive normalt kan förskjutas (vilket ofta är fallet), så kommer man dock att finna att man förlorar i specificitet när man försöker öka sensitiviteten och vice versa. Sensitivitet och specificitet beskriver sammantaget graden av överlappning mellan friska och sjuka för det aktuella testet och båda är nödvändiga för att förstå testets förmåga att skilja sjuka och friska åt.

- I kliniska sammanhang kan det vara användbart att tänka sig vad som händer om man har ett perfekt sensitivt test (se Faktaruta 2). Följden blir att alla som är testnegativa inte har sjukdomen, men man kan fortfarande inte vara riktigt säker för hur fördelningen är bland de testpositiva. Ett test med mycket hög sensitivitet är således bra för att utesluta sjukdom, vilket först kan verka kontraintuitivt. Det motsatta gäller för det perfekta specificika testet (Faktaruta 2). Man kan då vara

säker på att de som har ett positivt test verkligen är sjuka. Ett test med mycket hög specificitet är alltså bra för att bekräfta att sjukdom föreligger.

Om t ex situationen vid djup ventrombos är sådan att man med ett enklare screeningtest vill frikänna personer från djup ventrombos på ett någorlunda säkert sätt för att kunna avstå från vidare utredning och behandling, önskar man sig framför allt ett test med hög sensitivitet. Om man däremot vill selektera patienter för t ex en mer drastisk intervention som potentiellt också kan ha allvarliga biverkningar, som trombolysbehandling så vill man kanske i slutet av den diagnostiska strategin ha ett höggradigt specifikt test för att bekräfta att en trombos föreligger.

Positivt och negativt prediktionsvärde

Positivt prediktionsvärde (PPV) beskriver sannolikheten att patienten har sjukdomen givet att testet är positivt. Det negativa prediktionsvärdet (NPV) beskriver sannolikheten att patienten inte har sjukdomen givet att testet är negativt (Faktaruta 1). PPV och NPV beskriver alltså testets egenskaper från klinikerns perspektiv bättre än sensitivitet och specificitet. PPV och NPV är beroende av testets sensitivitet och specificitet (Faktaruta 1), men också av prevalensen av sjukdomen i den undersökta populationen. Om vi i vårt exempel med ultraljud i Faktaruta 1 i stället tänker oss att författarna undersökt en patientpopulation där djup ventrombos är ungefär hälften så vanlig i den aktuella populationen (omkring 13 procent i stället för 35 procent) så kommer det positiva och negativa prediktionsvärdet att förändras. En sådan tänkt situation visas i Faktaruta 3. Ju ovanligare sjukdomen är desto mer sjunker PPV för en given sensitivitet och specificitet. Ju vanligare sjukdomen är, desto lägre blir NPV. PPV och NPV för ultraljud uträknade enligt Faktaruta 1 är därför överförbart till andra kliniska situationer bara om prevalensen är ungefär jämförbar. Dessutom kan – enligt ovan – typ av ”case mix” ha betydelse för sensitivitet och specificitet och därmed för PPV och NPV.

Det vore en fördel att ha ett kvantitativt mått på ett tests prestanda som sammanfattar sensitivitet och specificitet och som är oberoende av

prevalensen. Ett sätt att lösa detta är att definiera PPV och NPV på ett mer sofistikerat sätt med hjälp av Bayesianskt tänkande. Likelihood-kvoter (LR) är dock enklare att räkna ut (Faktaruta 1) och har en del praktiska fördelar [11].

Likelihood-kvot

Positiv och negativ likelihood-kvot (”likelihood ratio”; LR) definieras i Faktaruta 1. En positiv likelihood-kvot (LR+) beskriver sannolikheten att vara testpositiv om man har sjukdomen genom sannolikheten att ha ett positivt resultat om man är frisk. En negativ likelihood-kvot (LR–) beskriver sannolikheten för ett negativt testresultat om man är sjuk dividerat med sannolikheten för ett negativt testresultat om man är frisk. En positiv likelihood-kvot kan alltså beskrivas som sensitivitet dividerat med (1-specificiteten) och vice versa kan en negativ likelihood-kvot definieras som (1-sensitivitet) dividerat med specificiteten. Fördelarna med dessa mått är flera:

- Om man vänjer sig vid begreppet LR och hur olika nivåer av LR påverkar sannolikheten att en testpositiv patient har sjukdomen (eller en testnegativ är frisk), kan man med ett enkelt kvantitativt mått få en känsla för hur värdefullt det diagnostiska testet är (Faktaruta 4). Om man har en uppfattning om prevalensen av sjukdomen i den patientgrupp som undersökes kan man med hjälp av formler och ett nomogram t o m kvantifiera sannolikheten att patienten har sjukdomen respektive är frisk givet ett positivt eller negativt test-resultat [2,4]. En sådan uträkning exemplifieras i Faktaruta 4 och 5.

- Det har utvecklats metoder för att använda LR i systematiska översikter och metaanlyser av egenskaperna hos diagnostiska test [1].

- LR är användbara när man studerar ett test som kan ha flera olika alternativ för var man drar gränsen för positivt respektive negativt test. Man kan visserligen för exempelvis patientpopulationen ange sensitivitet och specificitet om man skulle dela upp fynden i flera kategorier än två. Fynd vid lungskintigrafi vid misstänkt lungembolism uppdelas t ex ofta i de med hög, medelhög eller låg sannolikhet för emboli eller

i normala [4]. Ju fler sådana nivåer som införs desto mer svåröverskådligt och svårgripbart blir emellertid angivande av sensitivitet och specificitet för varje nivå. Om man i stället räknar ut ett LR för var och en av nivåerna, så blir det mer överblickbart av var man bör lägga gränsen i olika kliniska situationer.

Receiver operating characteristics

”Receiver operating characteristics” (ROC) är ytterligare ett sätt att sammanfatta testresultat när man har flera olika möjliga nivåer av gränsdragning mellan normalt och sjukt. Detta är användbart när man studerar ett laboratorietest med utfall längs en kontinuerlig skala. Av tradition från andra områden (t ex forskning om mottagande av radio- och radarsignaler) har man framställt ROC som en funktion där sensitivitet (sant positiva) är beroende av ett minusspecificitet (falskt positiva) för varje studerad gräns mellan sjukt och friskt, vilket illustreras med ett exempel i Faktaruta 6.

Ett test som ger en linje på eller mycket nära diagonalen i diagrammet illustrerar ett test som inte har något diskriminativt värde alls, medan ett test som har en kurva med ett vertex som ligger nära det vänstra övre hörnet har bra egenskaper. Genom en sådan grafisk framställning kan man bidra till beslut om var man på ett tests ROC-kurva tycker att den lämpligaste gränsen för normalt respektive onormalt ligger.

Rent matematiskt är den optimala gränsdragningen mellan sjukt och friskt i den punkt som ligger närmast diagrammets övre vänstra hörn. Det är dock inte säkert att detta kliniskt är det bästa valet. I olika sammanhang kan man välja att prioritera antingen sensitivitet eller specificitet. Man kan också jämföra olika test (även med statistisk prövning), med olika ROC-kurvor. ”Receiver operating characteristics” kan precis som LR användas i metaanalyser av test [1,3,8].

Test som inbegriper en tolkning, t ex avläsning av en röntgenbild, ger också en utvärdering med gränsdragningar längs en glidande skala. Olika röntgenologer med samma skicklighet kan välja att lägga sig på olika delar av en och samma ROC-kurva för att t ex fastställa på

datortomografibild om tecken till lungemboli föreligger eller inte. Om röntgenologerna sinsemellan har olika grad av skicklighet kompliceras bilden av att de då dessutom ligger sinsemellan på olika ROC-kurvor.

Problem i evaluering av diagnostiska test

Det finns flera metodologiska problem i utvärderingen av diagnostiska test, som är värda att tänka på vid studium av litteraturen och vid forskningsplanering. En del av problemen har kortfattat omnämnts ovan, men summeras här. De tre första punkterna är viktigast och utgör centrala bedömningsgraden för kvaliten i studier av diagnostiska test [9,10]:

- Bias kan införas om det nya testet och referenstestet inte utvärderas oberoende av varandra. En icke-blindad jämförelse kan leda både till över- och undervärdering av det nya testets diagnostiska egenskaper, oftast det förra. Oberoende bedömning av både det nya testet och referenstestet är en central del av designen i en studie av ett nytt test.

- Som nämndes ovan i avsnittet om sensitivetet och specificitet kan kliniska karakteristika hos de sjuka ha betydelse för beräkningen av sensitiviteten och karakteristika hos de friska patienterna ha betydelse för specificiteten. Den i studien ingående populationen av både sjuka och friska måste vara i rimlig grad representativ för hur problemet ser ut i klinisk vardag. Selektionsprocesser som ökar andelen av speciella patientgrupper gör att resultaten är svåra att generalisera. Om undersökarna inkluderat studiepersonerna konsekutivt minskar risken för selektion inom den eller de ingående institutionerna. Institutionerna i sig kan ju dock vara specialist- eller privatkliniker med ett utvalt klientel, så det är ändå viktigt att tänka igenom hur patientflödet skett.

- En viktig fråga inför läsningen av en studie är om alla patienter som genomgick det nya testet också har genomgått referenstestet. Om så inte är fallet, så måste man fråga sig varför så skedde och hur sådana patienter behandlats i analysen. Ibland har patienter som har vissa

karakteristika – t ex samtidiga kliniska symtom – genomgått referenstestet oftare. Då koncentreras oftast patienter med allvarliga sjukdomar i sjuka gruppen, varför intrycket av sensitivitet kan bli för högt. Av samma anledning sjunker specificiteten. I andra sammanhang betraktar man patienter som genomgått det nya testet, men inte undergått referenstestet som friska (detta är t ex vanligt i screeningstudier) och detta kan leda till en överskattning av både sensitivitet och specificitet.

- I många studiesammanhang förekommer att vissa testutfall inte är tolkningsbara. Detta kan påverka skattningen av både sensitivitet och specificitet om anledningen till att testet inte är tolkningsbart är relaterat till sjukdomsrisk. Ett exempel är om benödem vore en riskfaktor för bentrombos, samtidigt som ultraljud av benets vener ofta blir otolkningsbart vid benödem. Ett minimikrav är att icke tolkningsbara test redovisas och att författarna beskriver hur man hanterat dessa i analyserna.

- Ofta är den standard som ett test utvärderas mot (t ex när ultraljud jämförs med flebografi i exemplet i Faktaruta 1) ett annat test som inte heller det avslöjar ”sanningen”, se vidare Kapitel 1.7. Egenskaperna hos detta test måste tas hänsyn till. Om det nya testet och standardtestet samvarierar i många av sina egenskaper, så kan intrycket av sensitivitet och specificitet bli falskt höga.

- Precis som i andra kliniska studier, är beräkningarna i en studie skattningar av de sanna värdena. Medan detta har blivit mer och mer en självklarhet för mått som t ex överlevnad, så diskuteras inte detta så ofta för sensitivitet och specificitet. Man bör alltså ta hänsyn till att det alltid finns ett konfidensintervall kring dessa skattningar.

- Enligt diskussionen i avsnittet om ROC-kurvor ovan, så blir validiteten i utvärderingen svårtolkad om olika granskare vid test som inbegriper subjektiva bedömningar lägger sig på olika delar av en och samma ROC-kurva. Resultaten blir ännu mer svåröverskådliga om olika bedömare har olika grad av skicklighet och erfarenhet och därför

har olika ROC-kurvor. Tolkningen underlättas om det i artikeln finns information om hur samstämmiga olika utvärderare är, ofta beskrivet med ett så kallat kappavärde.

Givetvis finns andra metodologiska krav som gäller alla typer av studier. Några av de viktigaste är:

- rimlig statistisk styrka
- tydlig redovisning av tekniska metoder
- klar beskrivning av urvals- och uteslutningskriterier och
- klara definitioner av effektmått (”end-points”).

I många studier av diagnosmetoder saknas en bra beskrivning av de tekniker som använts. Detta är problematiskt i de situationer som detaljvariationen i teknik (t ex tidpunkten för kontrastinjektion och bildtagning) kan påverka sensitivitet och specificitet.

Tyvärr har studier av diagnostiska test ofta metodologiska brister, trots att de är viktiga för beslut om ofta stora investeringar av apparatur och införande av både dyra och tidskrävande metoder [9]. Dilemmat accentueras om metoden dessutom kan innebära obehagliga sidoeffekter för patienten. Det finns misstankar om att metodologiskt sämre studier systematiskt övervärderar testets prestanda [1,6].

I Appendix redovisas den checklista och mall som gruppen använt för granskning av diagnostiska studier samt dataextraktion från dessa.

Metaanalyser av diagnostiska test

Metaanalyser av diagnostiska test har börjat bli mer frekventa men är fortfarande förhållandevis ovanliga. Förutom att intresset inte varit lika stort som för behandlingsstudier, så är en av de viktigaste anledningarna att tillgängliga originalartiklar har så skiftande kvalitet att kvantitativa summeringar har tveksam validitet. I denna rapport har vi bara vid två tillfällen gjort en formell metaanalys och i det ena fallet med vissa reservationer.

För metaanalyser av diagnostiska test gäller som inom andra områden vissa grundläggande krav. Det måste finnas en klar frågeställning och en väl genomförd litteratursökning, vars metod tydligt redovisas. Inklusions- och exklusionskriterier för de ingående studierna och hur bedömningen gjordes ska framgå. Helst bör bedömningen av artiklarna ha gjorts av två eller flera bedömare oberoende av varandra. För diagnostiska studier gäller dessutom några speciella punkter:

- kvantitativa metoder ska ta hänsyn till hur sensitivitet och specificitet beroende av varandra
- särskilda metoder behövs om det är ett test med flera olika nivåer
- metaanalysen bör diskutera om det finns tecken till att skattningen av testets prestanda varierar med studiernas kvalitet
- slutligen bör man adressera problemet hur karakteristika hos de ingående studiepopulationerna påverkar generaliserbarheten till olika kliniska situationer.

Faktaruta 1

	Flebografi			
		DVT	Normal	Totalt
Ultraljud	DVT	70 (a)	1 (b)	71
	Normal	7 (c)	142 (d)	149
	Totalt	77	143	220

Sensitivitet	70/77=0,91	a/(a+c)
Specificitet	142/143=0,99	d/(b+d)
Positivt prediktionsvärde (PPV)	70/71=0,99	a/(a+b)
Negativt prediktionsvärde (NPV)	142/149=0,95	d/(c+d)
Prevalens av DVT	77/220=0,35; 35%	(a+c)/(a+b+c+d)
Positivt likelihood-kvot (LR+)	$\frac{70/77}{1/143}$ eller $\frac{0,9}{(1–0,99)}$	=91,0 [sensitivitet/(1-specificitet)] för tolkning av ett positivt testutfall
Negativt likelihood-kvot (LR–)	$\frac{7/77}{142/143}$ eller $\frac{(1–0,92)}{0,99}$	=0,09 [(1-sensitivitet)/specificitet] för tolkning av ett negativt testutfall

Faktaruta 2

		Flebografi	
		DVT	Normal
Ultraljud	DVT	77	1
	Normal	0	142

Det perfekt sensitiva testet. Man kan vara säker på att patienter som har ett normalt ultraljudsfynd inte har trombos.

		Flebografi	
		DVT	Normal
Ultraljud	DVT	70	0
	Normal	7	143

Det perfekt specifika testet. Man kan vara säker på att patienter som har ett positivt ultraljud har trombos.

Faktaruta 3

		Flebografi		
		DVT	Normal	Totalt
Ultraljud	DVT	20	1	21
	Normal	2	142	144
	Totalt	22	143	165

Prevalens 22/165 = 0,13 (13 procent)

PPV = 20/21 = 0,95

NPV = 142/144 = 0,99

Jämför med PPV och NPV i Faktaruta 1 där prevalensen var 0,35 men här i Faktaruta 3 är den 0,13. Skillnaderna i PPV och NPV blir mer drastiska om prevalensen ändras ytterligare. I en screeningsituation är prevalensen av sjukdomen oftast mycket låg vilket leder till ett lågt PPV.

Faktaruta 4

Likelihood-kvot (LR) (negativt LR är <1,0, positivt LR är >1,0)	Förändring i p(D)* efter test
>10 eller <0,1	Stor eller mycket stor
5–10 eller 0,1-0,2	Måttlig
2–5 eller 0,2-0,5	Liten, kan vara betydelsefull
1–2 eller 0,5-1	Liten och sällan betydelsefull

Pretest odds ** = är $\frac{\text{prevalens}}{(1 - \text{prevalens})}$

Post-test odds** = pretest odds x LR

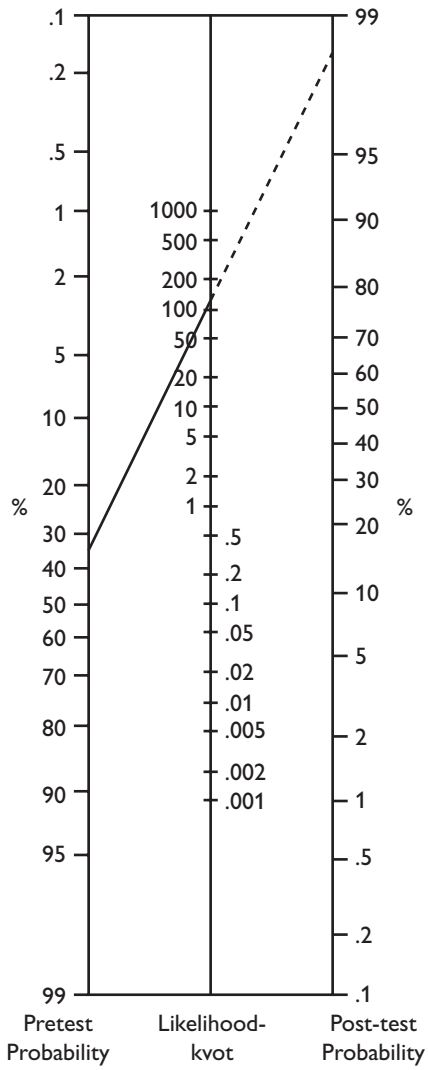
Post-test p(D) = $\frac{\text{post-test odds}}{(\text{post-test odds} + 1)}$

* sannolikheten (p) att patienten har sjukdomen (D)

** odds är $\frac{p(D)}{1 - p(D)}$ av sjukdomen.

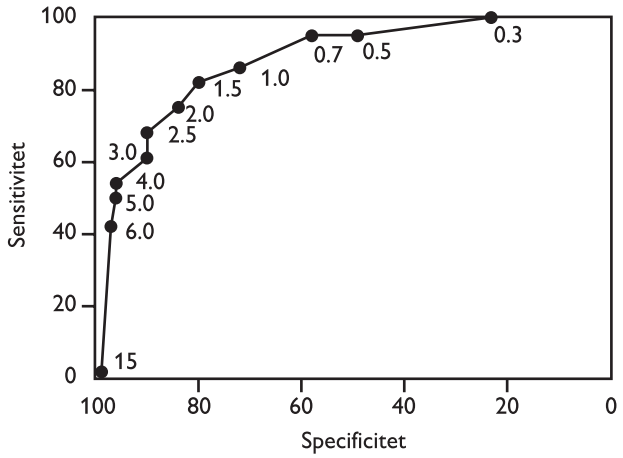
Om prevalensen är 50 procent är odds 50/50 och alltså 1. I exemplet i Faktaruta 1 är pretest odds 0,35/0,65. Post-test odds blir 0,54 x 91,0 = 49,0 och post-test p(D) alltså:
49,0/(49,0 + 1) = 0,98. De ultraljudspositiva har alltså teoretiskt i 98 procent av fallen en trombos.

Faktaruta 5



Textrutan illustrerar hur Fagans nomogram används för exemplet i Faktaruta 1: Sannolikhet före test (pretest probability) är 35 procent och likelihood-kvoten för ett positivt test är 91. Punkterna på den första och andra stapeln från vänster är således kända och linjen kan dras ut till sannolikhet efter positivt testutfall (post-test probability) på 98 procent. Med hjälp av nomogrammet slipper man således räkna med odds. Man kan förstås även räkna med negativa likelihood-kvoter (LR <1).

Faktaruta 6



ROC-analys av testegenskaperna hos en mikro-latexanalys för D-dimer. Referenstestet var flebografi. Siffrorna längs kurvan motsvarar olika cut-off för D-dimer-koncentrationer [7]. Just i detta exempel har författarna givit specificiteten på x-axeln. Det motsvarar emellertid precis det som man annars vanligen ser, nämligen att man ger (1-specificitet) från 0 i origo upp till 100 procent.

Referenser

1. Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. BMJ 2001; 323:157-62.

2. Fagan TJ. Letter: Nomogram for Bayes theorem. N Engl J Med 1975;293:257.

3. Irwig L, Tosteson AN, Gatsonis C, et al. Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 1994; 120:667-76.

4. 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.

5. Lensing AW, Prandoni P, Brandjes D, et al. Detection of deep-vein thrombosis by real-time B-mode ultrasonography. N Engl J Med 1989;320:342-5.

6. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061-6.

7. Lindahl TL, Lundahl TH, Fransson SG. Evaluation of an automated micro-latex D-dimer assay (Tina-quant on Hitachi 911 analyser) in symptomatic outpatients with suspected DVT. Thromb Haemost 1999; 82:1772-3.

8. Midgette AS, Stukel TA, Littenberg B. A meta-analytic method for summarizing diagnostic test performances: receiver-operating-characteristic-summary point estimates. Med Decis Making 1993;13: 253-7.

9. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA 1995;274:645-51.

10. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. Evidence-based medicine: how to practice and teach EBM. London: Churchill Livingstone; 1997.

11. Taube A, Malmquist J. [Count on your beliefs. Bayes' theorem in diagnosis]. Läkartidningen 2001;98:2910-3.