Diagnostikstudier bör följa internationella riktlinjer

Planering och utvärdering kräver teoretiska och praktiska överväganden

JONAS BJÖRK, docent, epidemiolog, Lunds universitet; FoU-centrum Skåne, Skånes universitetssjukhus, Lund
MIKAEL HELLSTRÖM, professor, överläkare, avdelningen för diagnostisk radiologi, Sahlgrenska akademin, Göteborgs universitet
INGEGERD MEJARE, professor emerita, projektledare, SBU (Statens beredning för medicinsk utvärdering), Stockholm
ULF NYMAN, docent, överläkare, Lunds universitet; Röntgen Öst, Centralsjukhuset Kristianstad
ulf.nyman@skane.se

I diagnostikstudier undersöks tillförlitligheten hos ett eller flera test vid identifiering av sjukdom eller värdering av patientens tillstånd. Ett diagnostiskt test kan omfatta kliniska fynd, biokemiska analyser, funktionsanalyser, bildgivande metoder och analys av vävnadsprov. Studier inom diagnostikens område kan vara behäftade med metodologiska svagheter, som gör att den diagnostiska förmågan eller patientnyttan hos indextestet felbedöms eller till och med överskattas. Det är därför viktigt med en transparent redovisning så att resultatens giltighet och generaliserbarhet går att bedöma.

Syftet med denna artikel är att beskriva vad man bör tänka på när man planerar, genomför, rapporterar eller utvärderar diagnostikforskning. Våra rekommendationer följer till stor del internationella riktlinjer, de sk STARD-kriterierna (Standards for the Reporting of Diagnostic accuracy studies; ‹http://www.stard-statement.org›) [1]. Liknande riktlinjer – QUADAS (Quality Assessment Tool for Diagnostic Accuracy Studies) [2, 3] och SBU:s metodbok ‹http://www.sbu.se/metodbok› – finns för utvärdering av diagnostikstudier i exempelvis en systematisk litteraturöversikt.

Grundläggande terminologi

Diagnostiska test ger ofta kvantitativa testsvar, som sedan kategoriseras, tex D-dimertest som används för att utesluta lungembolism och NT-proBNP (N-terminal brain natriuretic peptide) som används för att klassificera graden av hjärtsvikt. För att kunna fatta välgrundade beslut om vidare utredning eller behandling är det viktigt att känna till den diagnostiska tillförlitligheten (diagnostic accuracy) hos testet i den kliniska situationen. Test som utvärderas benämns indextest. Som jämförelse används ett referenstest, dvs ett test som bedöms som tillräckligt bra för att utgöra facit kring patientens verkliga tillstånd. Referenstestet kan i praktiken bestå av flera test och ibland också innebära att patienten följs upp över tid (Fakta 1).

Tillförlitligheten hos indextestet, dvs förmågan att korrekt identifiera, utesluta eller klassificera sjukdom i en specifik klinisk situation, beskrivs ofta med hjälp av sensitivitet, specificitet och prediktiva värden (Fakta 1) och ibland också med sannolikhetskvoter (likelihood ratios) [4, 5]. I situationer då det kvantitativa testsvaret används för att värdera patientens tillstånd utan klassificering kan överensstämmelsen med referensmetoden i stället illustreras med hjälp av sk Bland–Altman-diagram (Figur 1) och sammanfattas med kvantitativa mått på systematiskt fel (bias), precision och noggrannhet [6].

Klinisk situation och population avgör generaliserbarhet

Det är viktigt att precisera syftet med diagnostikforskningen,

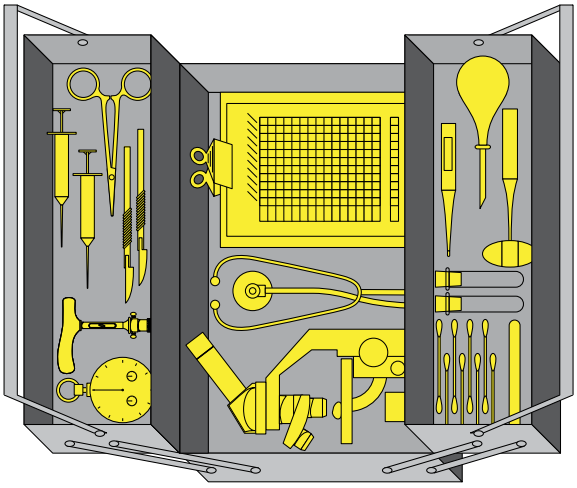


Illustration: Jakob Robertsson/Typoform

»Verktyg för klinisk forskning« är en artikelserie som omfattar 14 artiklar om grundläggande principer för hur man planerar och genomför kliniska forskningsstudier. Serien startade i nr 3/2013.

att ange vilken klinisk situation, vilken population och vilka undergrupper som ska studeras. Det finns en hierarki i utvärderingen av diagnostiska test, där inledande studier av ett nytt diagnostiskt verktyg (indextest) kan syfta till att avgöra om testet över huvud taget har förmåga att skilja ut – diskriminera – sjuka individer från friska. Diskrimineringsförmågan hos indextestet kan illustreras med en ROC-kurva (ROC = receiver operating characteristic) [5, 7] och användas för att hitta tröskelvärden hos kvantitativa testsvar för att identifiera eller utesluta sjukdom (Figur 2).

Fortsatta studier enligt den diagnostiska testhierarkin kan jämföra den diagnostiska förmågan (t ex sensitivitet, specificitet och prediktiva värden) med andra indextest eller bedöma hur mycket tillförlitligheten i diagnostiken förbättras om det nya testet kombineras med sådana test som redan används (clinical validity tests) [8, 9]. Ett nytt statistiskt mått, NRI (net reclassification improvement), kan användas för att beskriva

SAMMANFATTAT

Planering, genomförande, rapportering och kritisk granskning av diagnostikstudier bör följa internationella riktlinjer, de sk STARD- och QUADAS-kriterierna, så att resultatens giltighet och generaliserbarhet går att bedöma.
Basala kvalitetskriterier omfattar bl a adekvat beskrivning av studiedesign, inklusions- och exklusionskriterier, patientkarakteristika, utförande av index- och referenstest inklusive eventuell blindning vid bedömning av testresultaten.
Resultatet av indextestet för klassificering av individer som friska eller sjuka ska redovisas i en korstabell, tillsammans med sensitivitet, specificitet och prediktiva värden samt konfidensintervall som beskriver den statistiska osäkerheten.
Jämförelser av diagnostisk tillförlitlighet mellan olika indextest ska underbyggas med adekvata statistiska metoder.
Utöver diagnostisk tillförlitlighet är det viktigt att också utvärdera patientnyttan, dvs hur förändrad diagnostik påverkar morbiditet och mortalitet, samt hälsoekonomiska aspekter.

FAKTA 1. Diagnostisk tillförlitlighet

I en multicenterstudie, PIOPED II (Prospective investigation of pulmonary embolism diagnosis), undersöktes tillförlitligheten hos datortomografi (DT) för att korrekt identifiera lungembolism. Som referenstest användes en kombination av metoder: lungskintigrafi kombinerad med klinisk sannolikhetsbedömning, pulmonalisangiografi och ultraljud av nedre extremiteternas vener samt klinisk uppföljning [16]. Resultatet redovisas i tabellen nedan.

	Referenstest		
Indextest	**Sjuk**	**Frisk**	**Totalt**
Positivt	$a_1 = 150$	$a_2 = 25$	175
Negativt	$b_1 = 31$	$b_2 = 567$	598
Totalt	181	592	773

Sensitivitet och specificitet beskriver indextestets förmåga att påvisa sjukdom hos sjuka individer respektive utesluta sjukdomen hos friska individer.

Sensitivitet = sannolikheten att en sjuk patient (enligt referenstestet) blir klassad som positiv av indextestet =

$$\frac{a_1}{a_1+b_1} = \frac{150}{181} \approx 0{,}83 = 83 \text{ procent}$$ (95 procents konfidensintervall; 78–88 procent)

Specificitet = sannolikheten att en frisk patient (enligt referenstestet) blir klassad som negativ av indextestet =

$$\frac{b_2}{a_2+b_2} = \frac{567}{592} \approx 0{,}96 = 96 \text{ procent}$$ (95 procents konfidensintervall; 94–98 procent)

Sensitivitet och specificitet hänger intimt samman. Om sensitiviteten ändras påverkas specificiteten och tvärtom (Figur 2). Dessa mått kan emellertid inte användas direkt i den kliniska situationen för att bedöma sannolikheten att en patient är sjuk (eller frisk) utifrån ett testresultat. För en sådan bedömning behöver man i stället beräkna det prediktiva värdet av testresultatet.

Positivt prediktivt värde (PPV) = sannolikheten att patienten är sjuk om indextestet är positivt =

$$\frac{a_1}{a_1+a_2} = \frac{150}{175} \approx 0{,}86 = 86 \text{ procent}$$

Negativt prediktivt värde (NPV) = sannolikheten att patienten är frisk om indextestet är negativt =

$$\frac{b_2}{b_1+b_2} = \frac{567}{598} \approx 0{,}95 = 95 \text{ procent}$$

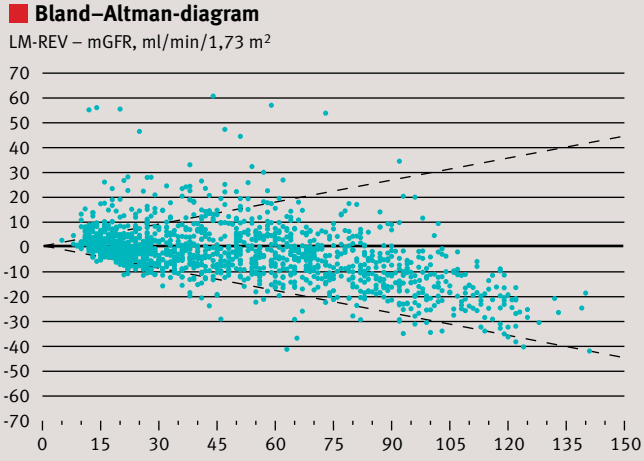
I den kliniska situation som undersöktes i PIOPED II-studien var prevalensen av lungembolism 23 procent, dvs innan indextestet (DT) utförs är den kliniska sannolikheten för sjukdom 23 procent. Vid positivt DT-fynd stiger den kliniska sannolikheten att patienten har lungemboli till 86 procent (PPV), medan den sjunker till 5 procent (1 – NPV) om DT-fyndet är negativt. Trots att DT visade normala fynd hos 17 procent av patienterna med lungemboli (sensitivitet 83 procent, dvs 17 procent falskt negativa) innebär således normalt DT-fynd i denna studie att lungemboli kan uteslutas med 95 procents säkerhet (NPV).

Uppföljningsstudier av patienter med normala DT-fynd visar i själva verket att risken för lungemboli kan uteslutas med 98–99 procents säkerhet [17]. Ett positiv DT-fynd innebär däremot en större osäkerhet med 14 procent falskt positiva (1 – PPV).

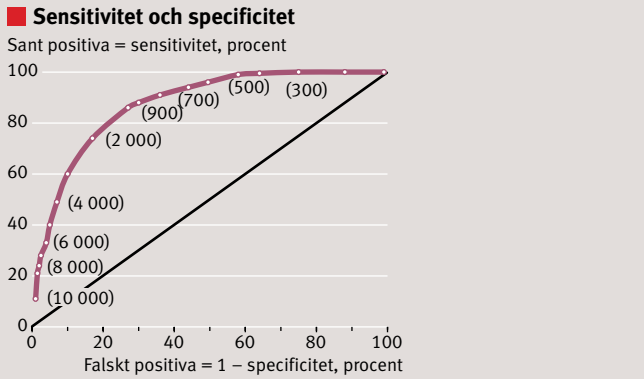
Som illustreras i Fakta 2 är prediktiva värden, men också sensitivitet och specificitet, starkt beroende av sjukdomens prevalens och allvarlighetsgrad i den kliniska situation där testet används.

hur mycket felklassificeringen minskar [10]. Diagnostikstudier kan också undersöka hur tillförlitligheten hos indextestet varierar i olika kliniska situationer och populationer.

Som en allmän tumregel är sensitiviteten högre och specifi-



Figur 1. Bland–Altman-diagram visar hur skillnaden mellan två kvantitativa mätmetoder (som ritas på y-axeln) beror av den verkliga nivån (som ritas på x-axeln) [18]. I figuren redovisas differensen mellan skattad njurfunktion (glomerulär filtrationshastighet [GFR] i ml/min/1,73 m²) utifrån den reviderade Lund–Malmö-formeln (LM-REV) och referenstestet, uppmätt GFR (mGFR) med plasmaclearance av johexol. Beräkningarna bygger på en svensk studie med 1 397 mätningar från patienter som remitterats för mätning av njurfunktion [19]. Figuren visar att indextestet (LM-REV) i den aktuella studien överskattar njurfunktionen vid låga uppmätta GFR-värden och underskattar njurfunktionen vid höga GFR-värden. De streckade linjerna visar gränserna för noggrannhet definierade som skattningar enligt indextestet, som avviker högst 30 procent från uppmätt GFR enligt referenstestet [6]. Om den verkliga nivån är okänd, dvs om referenstest saknas, uppskattas den verkliga nivån på x-axeln som medelvärdet av de båda metoderna som jämförs [20].



Figur 2. Sant positiva (sensitivitet) och falskt positiva (1 – specificitet) vid diagnostik av lungembolism utifrån olika tröskelvärden hos ett D-dimertest illustrerat i en ROC-kurva (ROC = receiver operating characteristic) [7] ritad med hjälp av publicerade data [21]. Siffrorna inom parentes anger olika D-dimernivåer i µg/l. Låga värden på D-dimertestet används för att utesluta lungembolism. Vid tröskelvärdet 500 µg/l är andelen sant positiva 99 procent (sensitivitet) och andelen falskt positiva 58 procent, dvs specificiteten är 42 procent. I studien är prevalensen 29 procent, och på motsvarande sätt som i Fakta 1 kan man räkna ut att lungembolism kan uteslutas med ≥99 procents säkerhet (negativt prediktivt värde) om D-dimervärdet är <500 µg/l. Den låga specificiteten vid denna gräns ger lågt positivt prediktivt värde (41 procent), dvs ett positivt D-dimertest beror på andra orsaker än lungembolism i 59 procent av fallen och är således ospecifikt. ROC-kurvan visar att specificiteten inte kan höjas utan att sensitiviteten samtidigt sjunker väsentligt. Om man i stället använder 700 µg/l som tröskelvärde sjunker sensitiviteten till 94 procent, specificiteten höjs till 56 procent och det negativa prediktiva värdet sjunker till 96 procent.

FAKTA 2. Variation hos sensitivitet, specificitet och prediktiva värden i olika kliniska situationer

En svensk metodstudie har undersökt den diagnostiska förmågan hos MDRD-formeln (Modification of diet in renal disease study), som används för att skatta njurfunktionen utifrån plasmakoncentrationen av kreatinin, ålder och kön i olika populationer med olika förekomst och grad av njursjukdom [22].

Resultaten (som redovisas i tabellen nedan) visar att sensitiviteten och specificiteten att upptäcka sänkt njurfunktion (glomerulär filtrationshastighet [GFR] <60 ml/min/1,73 m^2) varierar mellan 82 procent och 97 procent respektive 67 procent och 93 procent för olika patientgrupper.

Sensitiviteten är högst (och specificiteten lägst) i en population av njursjuka, medan sensitiviteten är lägst (och specificiteten högst) i en screeningundersökning av i huvudsak friska individer.

Lägg också märke till hur stor inverkan skillnaderna i prevalens, sensitivitet och specificitet i de olika populationerna har på de positiva och negativa prediktiva värdena (PPV och NPV).

PPV = sannolikheten att en patient som har skattad GFR <60 ml/min/1,73 m^2 enligt MDRD-formeln också har sänkt njurfunktion (uppmätt GFR <60 ml/min/1,73 m^2) enligt referenstestet.
NPV = sannolikheten att en patient som har skattad GFR ≥60 ml/min/1,73 m^2 enligt MDRD-formeln inte heller har sänkt njurfunktion (uppmätt GFR <60 ml/min/1,73 m^2) enligt referenstestet.

Population	Prevalens av sänkt njurfunktion, procent	Sensitivitet, procent	Specificitet, procent	PPV, procent	NPV, procent
Kroniskt njursjuka	84	97	67	94	81
Patienter remitterade för utredning	55	91	89	91	89
Hypotetisk screeningundersökning	4,5	82	93	36	99

citeten lägre ju högre prevalensen är av det tillstånd som ska diagnostiseras (Fakta 2). Resultaten av en studie som exempelvis genomförts bland patienter som remitterats för vidare utredning (hög prevalens = hög klinisk sannolikhet) kan därför inte utan vidare överföras till andra situationer, t ex rutinundersökningar med låg klinisk sannolikhet eller rena screeningsituationer.

På översta nivån i den diagnostiska testhierarkin undersöks om patienter som utsätts för indextestet får en behandling som leder till reducerad morbiditet och mortalitet eller om hälsoekonomiska analyser visar på förbättrad kostnadseffekt (clinical utility tests) [11]. Det är viktigt att man gör klart för sig på vilken nivå den aktuella diagnostikstudien befinner sig både när man genomför och när man utvärderar diagnostikforskning. Tyvärr saknas ofta studier som utvärderar den kliniska nyttan av att införa ett nytt diagnostiskt test.

Studieupplägg påverkar risk för snedvridning av resultat

Det är viktigt att studiens design och inklusions- och exklusionskriterierna för de testade patienterna beskrivs utförligt. Studiesyftet avgör vilket upplägg som är mest lämpligt. Flertalet diagnostikstudier är tvärsnittsundersökningar, dvs index- och referenstest utförs samtidigt och jämförs direkt utan ytterligare uppföljning. Prospektiv inklusion av konsekutiva patienter som alla genomgår både index- och referenstest oavsett utfall på de enskilda testen är normalt att föredra. Om det finns en fördröjning mellan utförandet av index- och referenstestet är det viktigt att detta redovisas tydligt så att det går att bedöma om det är rimligt att anta att patientens tillstånd inte ändrats under tiden.

Longitudinella undersökningar, dvs uppföljningar över tiden, kan syfta till att värdera patientnyttan av förbättrad diagnostik, t ex genom att studera om resultatet av indextestet kan förutsäga eller påverka sjukdomsförloppet om adekvat behandling ges. En longitudinell undersökning har i allmänhet högst evidensvärde om den är randomiserad, t ex en screeningundersökning där ett slumpmässigt urval av befolkningen erbjuds ett diagnostiskt test för att undersöka om tidig upptäckt av sjukdomen kan minska morbiditet eller mortalitet jämfört med övriga i befolkningen som inte erbjuds testet. Även välgjorda observationsstudier baserade på registerdata kan ha ett stort värde.

Studiens storlek ska motiveras med en s k styrkeberäkning (power-beräkning), som kan utföras med hjälp av konventionella metoder för binära utfall som finns implementerade i standardprogram för statistisk analys [12]. Om studien är en tvärsnittsundersökning som jämför olika diagnostiska test, bör den dimensioneras så att den har god chans (ofta uttryckt som minst 80 procents sannolikhet) att upptäcka kliniskt intressanta förbättringar i den diagnostiska tillförlitligheten, t ex ökad sensitivitet eller ökad specificitet.

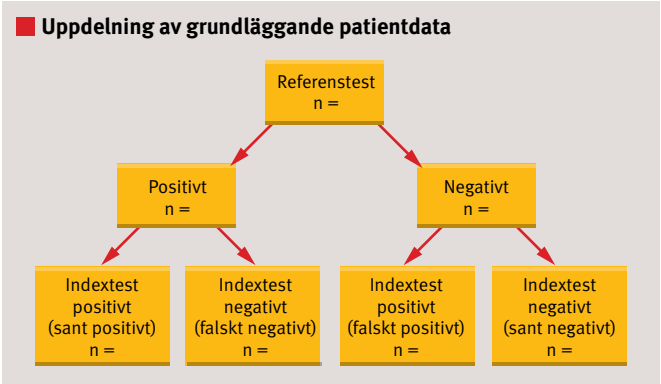
Antalet patienter som deltar i studiens olika faser bör redovisas i ett flödesdiagram så att bortfallets omfattning och eventuella konsekvenser går att bedöma (se Figur 1 i Bossuyt et al [1] som exempel). Selektionsfel (selection bias) uppstår om de inkluderade studiepatienterna skiljer sig systematiskt från den population som studien avsåg att utvärdera, vilket kan påverka såväl giltigheten som generaliserbarheten av resultaten. Risken för selektionsfel är särskilt överhängande i retrospektiva studier som baseras på redan insamlade patientdata från vårdsituationer där såväl index- som referenstest använts. I sådana patientmaterial kommer kliniska situationer där referenstestet sällan eller aldrig utförts att vara underrepresenterade, något som är viktigt att beakta när resultatens giltighet värderas.

I en nyligen publicerad SBU-rapport utvärderades formler för att skatta njurfunktionen baserade på plasmakoncentrationen av kreatinin eller cystatin C [13]. En tydlig brist i kunskapsunderlaget, som också påtalades i rapporten, var avsaknaden av prospektiva studier i populationer som sällan genomgår mätning av njurfunktionen med mera invasiva referenstest (såsom johexolclearance), t ex intensivvårdspatienter och andra patienter inlagda på sjukhus samt äldre multisjuka individer på vårdboende.

Selektionsfel kan också uppstå om referentestet bara utförs om indextestet är positivt, eftersom falskt negativa indexresultat då saknas i undersökningen. I ovan nämnda SBU-rapport exkluderades av denna anledning studier som endast omfattade patienter där indextestet (skattning av njurfunktionen med formel) indikerat sänkt njurfunktion. Selektionsfel kan också förekomma i prospektiva studier, exempelvis om hälsan hos deltagarna i en screeningundersökning redan vid studiestarten skiljer sig från hälsan i övriga befolkningen.

Genomförandet har stor betydelse för tillförlitligheten

Det är viktigt att index- och referenstest beskrivs på ett sådant sätt att noggrannheten i utförandet går att bedöma och så att det finns förutsättningar att värdera resultaten och eventuellt anamma den testade metoden i rutindiagnostiken. Precisionen hos kemiska analysmetoder som används för att ge kvantitativa provsvar redovisas normalt som variationskoefficienten (CV = coefficient of variation), dvs standardavvikelsen vid upprepade bestämningar av samma biologiska prov uttryckt i procent av medelvärdet. För indextest som ut-



Figur 3. Grundläggande demografiska, antropometriska och kliniska data för deltagarna i en diagnostikstudie bör, om antalen tillåter, redovisas uppdelade utifrån resultaten av både referens- och indextestet.

görs av kvalitativa bedömningsinstrument bör resultat redovisas, vilka anger testets tillförlitlighet vid upprepade bedömningar av samma tillstånd (intraobserver eller test–retest reliability) och för skilda bedömare (interobserver eller inter-rater reliability) [5].

Erfarenhet och kompetensnivå hos tolkare av testet bör beskrivas, eftersom det kan ha stor betydelse för utfallet. Blindning, dvs att resultatet av indextestet eller andra testresultat inte får påverka bedömningen av referenstestet och vice versa, är av stor vikt för att förhindra snedvridning av resultaten.

Valet av referenstest är mycket viktigt, eftersom brister i referenstestet kan leda till att tillförlitligheten hos indextestet underskattas och att testets sensitivitet och specificitet blir missvisande. Om referensmetoder med olika tillförlitlighet används, kan det vara svårt att jämföra resultaten mellan olika studier. Skillnader mellan olika referensmetoder kan exempelvis vara en viktig förklaring till de relativt stora skillnader i tillförlitlighet hos kreatininbaserade formler för skattning av njurfunktion som har rapporterats i olika populationer [14]. Brister i referenstestet kan också vara en förklaring till den relativt låga sensitiviteten hos datortomografi för att korrekt identifiera lungembolism, vilket redovisas i Fakta 1 [15].

Resultat – variation och statistisk osäkerhet

I diagnostikstudier bör grundläggande demografiska (t ex kön och ålder), antropometriska (t ex längd och vikt) och kliniska patientdata redovisas uppdelat utifrån resultatet av referenstestet (positivt/negativt, Figur 3). Ytterligare uppdelning av sådana bakgrundsdata utifrån resultatet av indextestet kan ge viktig information om vilka kliniska eller andra data som kan påverka den diagnostiska tillförlitligheten. Förekomst av andra diagnoser (samsjuklighet) och sjukdomens svårighetsgrad hos patienterna med positivt referenstest bör också redovisas, eftersom sådana faktorer kan ha stor betydelse för testets diagnostiska tillförlitlighet och generaliserbarhet.

Resultatet av indextestet för klassificering av individer som friska eller sjuka ska redovisas i en korstabell som i Fakta 1, tillsammans med sensitivitet, specificitet och prediktiva värden samt konfidensintervall som beskriver den statistiska osäkerheten. Jämförelser av diagnostisk tillförlitlighet mellan olika indextest som utförts på samma patienter kan också redovisas med konfidensintervall, och skillnader kan prövas med statistiska metoder för parade mätningar, t ex McNemars exakta test [5]. Om studien redovisar hur den diagnostiska tillförlitligheten varierar i olika undergrupper, kan statistiska metoder som värderar om den observerade variationen är mer än slumpmässig, t ex heterogenitetstest, med fördel användas.

Förutsättningar för att förbättra studiekvaliteten

Vi har i denna artikel diskuterat diagnostikstudier som värderar diagnostisk tillförlitlighet, men också beaktat studier av överensstämmelse mellan kvantitativa testsvar, longitudinella studier av patientnytta och randomiserade studier. Våra rekommendationer är i vissa fall mer långtgående än de internationella riktlinjer för diagnostikstudier som i dag finns att tillgå. Tillsammans skapar detta goda förutsättningar för att förbättra kvaliteten hos diagnostikstudier och bör vara känt av alla som utför, läser och granskar diagnostikstudier.

En transparent rapportering är visserligen ingen garanti mot fel i studieupplägg, urval av patienter, genomförande, statistisk analys eller tolkning, men den gör att sådana fel går att upptäcka och att konsekvenserna för resultatens giltighet och generaliserbarhet blir möjliga att bedöma.

Potentiella bindningar eller jävsförhållanden: Inga uppgivna.

LÄS MER Fullständig referenslista Läkartidningen.se

REFERENSER

1. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem. 2003;49:7-18.
2. Whiting P, Rutjes AW, Reitsma JB, et al. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol. 2003;3:25.
3. Reitsma JB, Moons KG, Bossuyt PM, et al. Systematic reviews of studies quantifying the accuracy of diagnostic tests and markers. Clin Chem. 2012;58:1534-45.
4. Brismar J, Jacobsson B. Definition of terms used to judge the efficacy of diagnostic tests: a graphic approach. AJR Am J Roentgenol. 1990;155:621-3.
5. Björk J. Praktisk statistik för medicin och hälsa. Stockholm: Liber AB; 2011.
6. Stevens LA, Zhang Y, Schmid CH. Evaluating the performance of equations for estimating glomerular filtration rate. J Nephrol. 2008; 21:797-807.
7. Brismar J. Understanding receiver-operating-characteristic curves: a graphic approach. AJR Am J Roentgenol. 1991;157:1119-21.
8. Linnet K, Bossuyt PM, Moons KG, et al. Quantifying the accuracy of a diagnostic test or marker. Clin Chem. 2012;58:1292-301.
9. Moons KG, de Groot JA, Linnet K, et al. Quantifying the added value of a diagnostic test or marker. Clin Chem. 2012;58:1408-17.
10. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. 2008;27:157-72; discussion 207-12.
11. Bossuyt PM, Reitsma JB, Linnet K, et al. Beyond diagnostic accuracy: the clinical utility of diagnostic tests. Clin Chem. 2012;58(12): 1636-43.
12. Dupont WD, Plummer WD Jr. Power and sample size calculations. A review and computer program. Control Clin Trials. 1990;11: 116-28.
13. Skattning av njurfunktion. Stockholm: SBU (Statens beredning för medicinsk utvärdering); 2012. Rapport 214.
16. Stein PD, Fowler SE, Goodman LR, et al. Multidetector computed tomography for acute pulmonary embolism. N Engl J Med. 2006; 354:2317-27.
17. Nyman U. Radiologisk diagnostik av akut lungembolism. I: Gottsäter A, Svensson PJ, redaktörer. Klinisk handläggning av venös tromboembolism. 1 ed. Lund: Studentlitteratur AB; 2010. p. 233-54.
18. Krouwer JS. Why Bland-Altman plots should use X, not (Y+X)/2 when X is a reference method. Stat Med. 2008;27:778-80.
19. Björk J, Jones I, Nyman U, et al. Validation of the Lund-Malmö, Chronic Kidney Disease Epidemiology (CKD-EPI) and Modification of Diet in Renal Disease (MDRD) equations to estimate glomerular filtration rate in a large Swedish clinical population. Scand J Urol Nephrol. 2012;46: 212-22.
20. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1:307-10.
21. Perrier A, Desmarais S, Goehring C, et al. D-dimer testing for suspected pulmonary embolism in outpatients. Am J Respir Crit Care Med. 1997;156:492-6.
22. Björk J, Grubb A, Nyman U. Variability in diagnostic accuracy can be estimated using simple population weighting. J Clin Epidemiol. 2009;62:54-7.