REVIDERAD 2017

7 Tillförlitlighet av tester och bedömningsmetoder

Inledning

Detta avsnitt handlar om utvärdering av *tester*, dels medicinska testmetoder och dels standardiserade bedömningsmetoder, och specifikt om testernas diagnostiska tillförlitlighet (eng. diagnostic accuracy). Tester kan inkludera kliniska fynd, symtom, social problematik, bilder och biokemiska analyser. De kan användas för flera olika ändamål, till exempel för att ställa diagnos, bedöma risk eller prognos, följa ett sjukdomsförlopp eller värdera effekten av behandling. En speciell form av diagnostik är screening. Screening för ett hälsoproblem görs på en del av befolkningen där prevalensen vanligtvis är mycket lägre än hos dem som sökt vård för ett misstänkt problem. Vid screening är därför möjligheterna att förutsäga vilka som är sjuka respektive friska sämre. Bedömningsmetoder kan omfatta standardiserade formulär för utredning och bedömning av en situation, funktion eller behov av hjälp inom socialtjänst och skola samt diagnoser av till exempel psykiatriska tillstånd.

Med diagnostisk tillförlitlighet avses hur väl ett test eller en bedömningsmetod kan särskilja dem som har ett visst tillstånd från dem som inte har det. Tillståndet kan såväl vara en diagnos som problem till exempel i familjen eller med droganvändning. Avsnittet är avgränsat till utvärderingar av den diagnostiska tillförlitligheten uttryckt som sensitivitet (känslighet) och specificitet (träffsäkerhet).

Utvärdering av diagnostisk tillförlitlighet följer samma principiella tillvägagångssätt som utvärderingar av interventioner. Det finns riktlinjer för genom-

förande och rapportering av studier om diagnostisk tillförlitlighet, Standards for Reporting of Diagnostic Accuracy, STARD [1].

Terminologi

Testet som ska utvärderas (indextest)

Det medicinska test eller den bedömningsmetod som ska utvärderas kallas indextest. De flesta studier fokuserar på att undersöka tillförlitligheten hos ett enstaka test. I praktiken behövs det ofta flera tester för att kunna ställa en diagnos eller avgöra om det till exempel finns ett socialt problem. Vid parallell testning görs alla tester på en gång. Om ett av testerna ger positivt utslag räcker det för att ställa en diagnos eller avgöra att individen har det problem som testet avser. Vid seriell testning ges testerna i följd och alla testen måste ge positivt utslag för att räcka för en diagnos eller bedömning. Så fort ett test ger negativt resultat avbryts testningen. Parallell testning med multipla tester ökar i allmänhet sensitiviteten, medan specificiteten sjunker, och andelen falskt positiva testresultat ökar. Seriell testning maximerar specificiteten, medan sensitiviteten sjunker, och andelen falskt negativa testresultat ökar [2].

Tester kan fylla olika funktioner i en utredning. Avsikten med indextestet kan vara att ersätta ett befintligt test. Ett exempel är mammografiscreening där befintligt test (bildgranskning utförd av två bröstradiologer) jämförs med nytt test (bildgranskning utförd av en bröstradiolog + datorstödd analys). Testet kan också vara ett tillägg till befintligt test. Slutligen kan testet användas som en första undersökning för att sortera bort individer från fortsatt testning, så kallat triage. Som exempel kan nämnas användning av Ottawa Ancle rules vid misstanke om frakturer i foten. Ottawa Ancle rules är en enkel klinisk undersökning av patienter med fot- och/eller ankelsmärtor och ger mycket låg andel falskt negativa fynd. Undersökningen används därför ofta för att reducera antalet onödiga röntgenundersökningar.

Indextestets roll i utredningen ska definieras i projektplanen. Det är också viktigt att känna till vilken annan information som finns tillgänglig för den undersökta populationen, innan tillförlitligheten utvärderas. Om till exempel studien endast rekryterar deltagare som haft positivt utslag på ett tidigare test är sannolikheten för att deltagarna har tillståndet högre än för en population som inte har gjort ett tidigare test. Det kommer att påverka det undersökta testets sensitivitet och specificitet.

Jämförelsetest (Referenstest)

Den diagnostiska tillförlitligheten hos indextestet bedöms genom att jämföra det med ett referenstest eller referensstandard (ibland benämnt gold standard). Referensstandarden ska representera det bästa tillgängliga sättet för att skilja mellan dem som har problemet ifråga och de som inte har det. Ett exempel är

histologisk undersökning efter biopsi av bröstvävnad för att fastställa närvaro/frånvaro av bröstcancer.

Ofta saknas dock en referensstandard som är tillräckligt bra. Alternativet är då att konstruera ett referenstest. Det kan göras på flera olika sätt (Faktaruta 7.1).

Faktaruta 7.1
Alternativa tillvägagångssätt när det saknas en referensstandard [3,4].

Sammansatt referensstandard. Här kombineras flera (var för sig otillräckliga) referenstester till ett sammansatt mått. Oftast räcker det med att ett av referenstesterna ger positivt utslag för att det ska räknas som närvaro av tillståndet. Ett exempel är en studie där man undersökte tillförlitligheten hos en antigenanalys för att diagnostisera Chlamydia trachomatis-infektion [5]. Två referenstester kombinerades: odling och DNA-analys (PCR, polymerase chain reaction). Om antingen odling eller PCR visade positivt resultat räknades det som bekräftad infektion. Om båda referenstesterna var negativa, var diagnosen inte infektion.

Panel- eller konsensusdiagnos. Här kombineras resultat av olika tester med till exempel kliniska karakteristika och prognostisk information, som tillsammans ger en pragmatisk validering av sjukdomen eller tillståndet. Referensstandarden valideras då med hjälp av en stor mängd empiriska data, ofta genom internationella konsensusprocedurer med expertpaneler eller med så kallad Delfi-procedur [4]. Ett exempel är DSM – kriterierna för psykiatriska tillstånd. Ett annat exempel är diastolisk hjärtsvikt där European Society of Cardiology rekommenderar att diagnosen baseras på symtom och kliniska fynd understött av EKG, röntgen, Dopplerekokardiogram och biomarkörer [6].

Validering vid uppföljning. En annan modell är att validera indextestet genom en prospektiv studiedesign, där observation enbart eller utfall av en behandling relateras till symtom och tester vid start. Detta kan betraktas som en fördröjd typ av verifiering. Här används ofta andra utfallsmått som behandlingsutfall och relativ risk. Ett exempel är formulär för att bedöma risken för att en patient ska göra ett suicidförsök. Här var referenstestet antalet individer som faktiskt hade genomfört ett suicidförsök vid uppföljning minst ett år senare. Ett annat exempel gäller tester till små förskolebarn för att förutsäga dyslexi. Referenstestet var huruvida barnet hade dyslexi några år senare då det var läskunnigt.

Statistiska modeller. Här kombineras klinisk information och andra testresultat i statistiska modeller, som genererar en sannolikhet för att till exempel hjärtsvikt föreligger [4].

Utfallsmått

De klassiska måtten för att beskriva diagnostisk tillförlitlighet är sensitivitet och specificitet. Med sensitivitet avses den andel av dem som har tillståndet som identifieras med indextestet (sant positiva). Specificitet definieras som den andel av dem som inte har tillståndet som identifieras med indextestet (sant negativa). Den diagnostiska tillförlitligheten representeras ofta i en så kallad fyrfältstabell

som visar antal individer är sant positiva, falskt positiva, sant negativa samt falskt negativa (Figur 7.1).

Figur 7.1
Utfallsmått för diagnostisk tillförlitlighet

	Har problemet enligt referensstandard	Har inte problemet enligt referensstandard
Har problemet enligt test som ska utvärderas	TP (sant positiv)	FP (falskt positiv)
Har inte problemet enligt test som ska utvärderas	FN (falskt negativ)	TN (sant negativ)

$$\text{Sensitivitet} = \frac{TP}{TP+FN} \qquad \text{Specificitet} = \frac{TN}{FP+TN}$$

Om prevalensen av tillståndet är känd kan andra mått beräknas, till exempel positivt och negativt prediktionsvärde (PPV respektive NPV), sannolikhetskvoter (likelihood ratio) och diagnostisk oddskvot (DOR).

Diagnostisk tillförlitlighet bygger alltså på antalet individer som har respektive inte har ett tillstånd, det vill säga värdena är dikotoma. Många tester ger dock kontinuerliga resultat, till exempel blodsockernivå eller vikt. För att kunna avgöra tillförlitligheten sätts ett tröskelvärde (eng. cut-off) där de som ligger över tröskelvärdet anses ha tillståndet. Tröskelvärdet är inte självklart. Oftast sätts tröskelvärdet där man bedömer att sannolikheten för att skilja mellan tillstånd/inte tillstånd är störst. Tröskelvärdet kan därmed variera mellan populationer med olika prevalens för tillståndet. Denna något godtyckliga gräns påverkar sensitiviteten och specificiteten. Om tröskelvärdet sänks blir effekten att det blir fler falskt positiva och färre falskt negativa. Om tröskelvärdet höjs blir resultatet det omvända.

"Tillräckligt god tillförlitlighet"

Tester ger sällan en hundraprocentigt säker diagnos eller bedömning men kan ge tillräcklig information för att fastställa eller utesluta ett tillstånd. Vad som ska anses vara tillräckligt god tillförlitlighet beror på i vilket sammanhang testet används. I vissa fall krävs en mycket hög diagnostisk tillförlitlighet, till exempel genetiska metoder för att bedöma kromosomavvikelser hos foster, medan lägre diagnostisk tillförlitlighet kan vara tillräcklig i andra fall. En diagnos kan till exempel vara tillräckligt säker för att den förväntade nyttan av att behandla patienten överväger de förväntade konsekvenserna av att inte behandla. Ofta tvingas man göra kompromisser och välja antingen hög sensitivitet eller hög specificitet. Då får man bedöma vad som får minst allvarliga konsekvenser: att behandla personer som inte har tillståndet eller att inte behandla sådana som har tillståndet. Avvägningen måste ta hänsyn till risk för biverkningar, behandlingskostnader och etiska aspekter.

Definiera syftet med översikten

Som vid alla systematiska översikter är en genomtänkt och tydlig formulering av frågan/frågorna väsentlig. För läkemedel finns en strängt reglerad internationell standard (hierarkisk fyrfas-modell), där vissa villkor måste vara uppfyllda i varje fas innan man får fortsätta till nästa (fas 0 är inledande, fas IV undersöker effekter och biverkningar på patienter på lång sikt efter marknadsgodkännande). För tester finns inte någon liknande hierarki som är allmänt vedertagen men däremot flera förslag [7]. En av dem, som ofta kallas den diagnostiska forskningens struktur [8] består av fyra nivåer av frågor (Faktaruta 7.2). Vilken eller vilka frågor som ska ingå i en översikt beror på det aktuella kunskapsläget. För SBU:s vidkommande gäller översikterna oftast steg 3 och 4.

Faktaruta 7.2
Den diagnostiska forskningens arkitektur.

Steg 1. Skiljer sig testresultaten hos individer som har det sökta problemet från testresultaten från dem som inte har det?
Steg 1-studier har ofta fall–kontrolldesign där en grupp individer som har problemet ifråga jämförs med en grupp som inte har det. Resultaten presenteras ofta som korrelation eller skillnader i medelvärden mellan grupperna. Ett positivt utfall i en Steg 1-studie gör det naturligt att gå vidare till nästa steg.

Steg 2. Är sannolikheten större att individer med ett visst testresultat har problemet ifråga jämfört med individer med andra testresultat?
Även här används oftast fall-kontrolldesign, dvs individerna antingen har eller saknar tillståndet. Resultaten presenteras som sensitivitet och specificitet.

Steg 3. Kan testresultat skilja ut individer med respektive utan problemet ifråga hos en grupp där det är rimligt att misstänka de har problemet?
Med steg 3-studier undersöks den diagnostiska tillförlitligheten i en population där testet är tänkt att användas. Individerna kan ha olika grad av tillståndet. Ofta används tvärsnittsstudier där samtliga deltagare undersöks med såväl index- som referenstest. Exempel 7.1 i slutet av avsnittet illustrerar hur Steg 1- och Steg 2-studier, som båda är fall–kontrollstudier, överskattar tillförlitligheten, och att Steg 3-studier är nödvändiga för att bestämma tillförlitligheten hos ett test i praktiken.

Steg 4. Ger testet någon effekt på hälsa och välbefinnande för dem som genomgår testet jämfört med liknande individer som inte genomgår testet?
Denna fråga avser det egentliga nyttan av testet för patienten eller klienten. Utfallet mäts i hälsoresultat som följd av om testresultatet påverkar handläggning eller val av behandling. Ibland kan nyttan vara uppenbar som till exempel en korrekt diagnos vid livshotande tillstånd. Relativt ofta handlar det emellertid om tester för tidig upptäckt av asymtomatisk sjukdom, till exempel PSA-prov (PSA = prostataspecifikt antigen) för tidig upptäckt av prostatacancer. Då kan Steg 4-frågan bara besvaras genom att följa patienter som randomiseras till det diagnostiska testet/alternativt test (eller inget test).

Ett annat sätt att dela in bedömningen av värdet av ett test eller en bedömningsmetod [9,10] är:

1. teknisk tillförlitlighet (analytisk precision och validitet)
2. diagnostisk tillförlitlighet (accuracy)
3. effekten på val av fortsatt handläggning eller behandling
4. effekten av behandlingen på patientens (klientens) välbefinnande
5. effekten på utfallet av behandlingen av patienten (klienten)
6. kostnad–nytta, kostnadseffektivitet.

Modellen kan vara användbar för att skilja mellan olika typer av studier, men den kan inte ses som en nödvändig sekvens, eftersom utvärderingen av tester sannolikt inte är linjär utan snarare cyklisk och repetitiv [7]. Separata systematiska översikter kan göras för vart och ett av dessa steg. Punkterna 3–5 i modellen motsvarar Steg 4-frågan, där värdet av ett test efterfrågas, det vill säga ”blir utfallet (resultat till följd av testutfallet) bättre för dem som genomgår testet jämfört med liknande patienter eller klienter som inte genomgår testet?”.

Formulera inklusions- och exklusionskriterier

I interventionsstudier används PICO (population, intervention, control, outcome) för att beskriva och sortera kriterier för inklusion respektive exklusion. För diagnostiska studier används PICO på motsvarande sätt (population, indextest, jämförelsetest (comparator), outcome (tillståndet referenstestet) (Exempel 7.1). Liksom för interventionsstudier lönar det sig att noga tänka igenom inklusions- och exklusionskriterierna. Det gäller till exempel vilken/vilka populationer som är relevanta, och vilken typ av studier som ska ingå i rapporten. Det kan också gälla vilka tester som ska undersökas, vilken/vilka referenstester som ska accepteras och längsta tid som får gå mellan administration av index- och referenstesterna.

Exempel 7.1
Inklusionskriterier formulerade som PICO [11].

PICO	Inklusionskriterier
Population	Patienter/klienter som kan förväntas få undersökningen eller testet i klinisk praxis/socialtjänst
Indextest	Det evaluerade indextestet
Comparator (jämförelsetest)	Eventuellt jämförelsetest
Utfall (outcome)	Tillståndet (referenstestet)

Bedömning av enskilda studier

Granskningen av studier omfattar dels huruvida det finns risk för systematiska fel (bias) och dels huruvida studien är tillämpbar, det vill säga en bedömning av intern och extern validitet. Den externa validiteten bestäms i hög grad av inklusionskriterierna för deltagarna, till exempel svårighetsgrad av tillståndet eller samsjuklighet. Är resultaten tillämpliga på den population(er) som översiktens frågeställning avser? En studie kan ha god metodologisk kvalitet, men relevansen kan ifrågasättas vid närmare bedömning av studien.

Risker för systematiska fel är delvis desamma som för interventionsstudier. Det gäller till exempel blindning. I interventionsstudier bör de som bedömer utfallet av en behandling vara blindade med avseende på vilken behandling patienten fått. I studier om diagnostisk tillförlitlighet bör bedömare vara blindade avseende utfall av indextestet då utfallet av referenstestet bedöms (och vice versa). De viktigaste källorna till bias i diagnostiska studier är sammanställda i Tabell 7.2. Som framgår av tabellen, finns en rad förhållanden, både i studiedesign och genomförande, som kan leda till systematiska fel eller variation. Det är dock osäkert hur mycket eventuella systematiska fel påverkar den beräknade diagnostiska tillförlitligheten och om de leder till över- eller underskattning [12].

Tabell 7.2
Källor till risk för systematiska fel i studier om diagnostisk tillförlitlighet [12,13].

Typ av bias	När inträffar det?	Under- eller överskattning av diagnostisk tillförlitlighet
Population		
Spektrumbias	När deltagarna inte representerar hela det spektrum av svårighetsgrad av problemet eller diagnosen som är relevant	Kan leda till såväl över- som underskattning
Selektionsbias	När deltagarna inte inkluderas konsekutivt eller slumpvis	Leder ofta till överskattning
Indextest		
Informationsbias (review bias)	När resultaten av indextestet tolkas med kännedom om resultatet från referenstestet	Leder ofta till överskattning. Om mindre information finns tillgänglig jämfört med i klinisk praxis, kan det leda till underskattning
Klinisk review bias	När data som ålder, kön och symtom finns tillgänglig då indextestet tolkas (gäller framför allt röntgenbilder)	Leder till högre sensitivitet men har liten påverkan på specificitet
Referensstandard		
Fel klassifikationsbias	När referensstandarden är bristfällig och inte alltid klassificerar deltagarna korrekt	Kan leda till såväl över- som underskattning
Partiell verifikationsbias	När enbart ett icke randomiserat urval av deltagare testas med referensstandard	Leder ofta till överskattning av sensitiviteten. Effekten på specificiteten varierar
Differentiell verifikationsbias	När vissa deltagare testas med en alternativ referensstandard, speciellt om selektionen till referenstest beror på resultatet av indextestet	Leder ofta till överskattning

Tabellen fortsätter på nästa sida

Tabell 7.2 fortsättning

Typ av bias	När inträffar det?	Under- eller överskattning av diagnostisk tillförlitlighet
Inkorporations-bias	När indextestet är en del av en (sammansatt) referensstandard	Leder ofta till överskattning
Sjukdoms-progressionsbias	När patientens tillstånd förändras mellan administrering av indextest och referensstandard	Över- eller underskattning
Informationsbias	När referensstandard tolkas med kännedom om resultatet av indextestet	Leder ofta till överskattning
Dataanalys		
Exkluderade data	När data som inte går att tolka och när bortfall av patienter inte inkluderas i analysen	Leder ofta till överskattning

Som nämndes i inledningen finns en internationell riktlinje, STAR D, för hur studier om diagnostisk tillförlitlighet ska genomföras och rapporteras [1]. Det finns också en checklista, QUADAS, med frågor som speglar komponenterna i STAR D [14]. QUADAS är i första hand anpassad för att bedöma tvärsnittsstudier. Nuvarande version av checklistan är QUADAS-2 som täcker både intern och extern validitet (Bilaga 4). Checklistan är uppdelad i fyra moduler: urval av deltagare, indextest, referenstest och ”flow and timing”. Den sistnämnda modulen tar upp hur lång tid som går mellan index- och referenstest samt huruvida samtliga deltagare gått igenom båda testerna. Varje modul omfattar frågor om både risk för systematiska fel (bias) och tillämplighet. Frågorna har tre svarsalternativ: ”ja”, ”nej” eller ”oklart”. Det finns tolkningsanvisningar för de olika alternativen.

QUADAS-2 ska ses som ett verktyg och hjälpmedel vid bedömning av risk för systematiska fel och tillämplighet. Vissa frågor i QUADAS-2 kräver ett visst mått av subjektiv värdering, medan andra är mer ”svart–vita”. För de mer subjektiva frågorna, som till exempel bedömning av om sammansättningen av deltagare är tillfredsställande, är det viktigt att formulera klara riktlinjer för hur studierna ska bedömas. Det kan göras genom att bedömarna, oberoende av varandra, gör en pilot-checklista på några studier och diskuterar eventuella skiljaktigheter. De slutgiltiga bedömningsgrunderna sammanställs i en manual.

Referenser

1. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ 2003;326:41-4.

2. Fletcher RH, Fletcher SW. Clinical epidemiology. The essentials. Philadelphia, Pennsylvania USA, Lippincott Williams & Wilkins; 2005.

3. Reitsma JB, Rutjes AW, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. J Clin Epidemiol 2009;62:797-806.

4. Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol Assess 2007;11:iii, ix-51.

5. Jang D, Sellors JW, Mahony JB, Pickard L, Chernesky MA. Effects of broadening the gold standard on the performance of a chemiluminometric immunoassay to detect Chlamydia trachomatis antigens in centrifuged first void urine and urethral swab samples from men. Sex Transm Dis 1992;19:315-9.

6. Paulus WJ, Tschope C, Sanderson JE, Rusconi C, Flachskampf FA, Rademakers FE, et al. How to diagnose diastolic heart failure: a consensus statement on the diagnosis of heart failure with normal left ventricular ejection fraction by the Heart Failure and Echocardiography Associations of the European Society of Cardiology. Eur Heart J 2007;28:2539-50.

7. Lijmer JG, Leeflang M, Bossuyt PM. Proposals for a phased evaluation of medical tests. Med Decis Making 2009;29:E13-21.

8. Sackett DL, Haynes RB. The architecture of diagnostic research. BMJ 2002;324:539-41.

9. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making 1991;11:88-94.

10. Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason. Science 1959;130:9-21.

11. SBU. Rotfyllning. En systematisk litteraturöversikt. In, Stockholm. Statens beredning för medicinsk utvärdering (SBU); 2010. SBU-rapport nr 203.

12. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004;140:189-202.

13. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM, Cochrane Diagnostic Test Accuracy Working G. Systematic reviews of diagnostic test accuracy. Ann Intern Med 2008;149:889-97.

14. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. 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.