

# Diagnostik: nye metodologiske initiativer

Ledende overlæge Erik Magid,  
ledende overlæge Per E. Jørgensen,  
forskningslæge Lone G.M. Jørgensen,  
reservelæge Jonna Skov Madsen,  
overlæge Torben Bjerregaard Larsen & lektor Jørgen Hilden\*

Amager Hospital, Klinisk Biokemisk Afdeling,  
Odense Universitetshospital, Afdeling KKA,  
Vejle Sygehus, Klinisk Biokemisk Afdeling og  
Københavns Universitet, Det Sundhedsvidenskabelige Fakultet,  
Biostatistisk Afdeling

Evidensgrundlaget for diagnostik bør forbedres. Det er konklusionen af en årrækkes internationale, kritiske analyser af diagnostiske studier. Kritikens hovedpunkter er:

- Mange publikationer om diagnostiske undersøgelers egnethed til at klassificere korrekt (diagnostisk træfsikkerhed) imødekommer ikke grundlæggende metodologiske krav. Dette kan føre til forkerte, oftest for optimistiske vurderinger af den pågældende test. Metodologiske mangler forringer også værdien af metaanalyser og systematiske *reviews* af den diagnostiske litteratur.
- Mange nye undersøgelser tages i brug uden metodologisk acceptable vurderinger af den pågældende undersøgelses konsekvenser for patienten (*outcome*).
- Kravene til de evalueringsprocesser, som nye undersøgelser bør gennemgå, før de tages i brug i klinisk praksis, bør tydeliggøres med henblik på at forebygge et overflødig forbrug og uhensigtsmæssige, i værste fald skadelige konsekvenser.

En aktuell undersøgelse af danske klaringsrapporter understøtter kritikken af det mangelfulde evidensgrundlag for brugen af diagnostiske undersøgelser [1].

Når det gælder terapi er arbejdet med at forbedre evidensgrundlaget længere fremme, såvel metodologisk som organisatorisk. Derfor har det været nærliggende at overføre nogle af de principper og metoder, som har vist sig at være effektive, fra det terapeutiske til det diagnostiske område. Det har ført til tiltag, som det er denne artikels formål at gøre opmærksom på: Standards for Reporting of Diagnostic accuracy (STARD)-initiativet og udgivelsen af en ny bog om evidensgrundlaget for klinisk diagnostik.

Forbilledet for STARD er Consolidated Standards Of Reporting Trials (CONSORT). CONSORT har vundet bred tilslutning og har ført til forbedret tilrettelæggelse og rapportering af randomiserede, kontrollerede undersøgelser af terapeutiske interventioner [2]. STARD-gruppen har nu fremlagt to artikler, som i begyndelsen af 2003 blev publiceret i ind-

flydelsesrige medicinske tidsskrifter, deriblandt BMJ, JAMA, Annals of Internal Medicine, Radiology og Clinical Chemistry [3, 4]. De to artikler er frit tilgængelige på <http://www.clinchem.org/cgi/content/full/49/1/1> og <http://www.clinchem.org/cgi/content/full/49/1/7>. STARD-gruppens hjemmeside findes via »Further initiatives« på <http://www.consort-statement.org/>.

Bogen: »The evidence base of clinical diagnosis«, som udkom i 2002, er redigeret af JA Knottnerus, der har samlet en kreds af bidragydere med tilknytning til bevægelserne *evidence-based medicine* og Cochrane-samarbejdet [5].

## STARD-initiativet: forbedret rapportering om diagnostiske undersøgelser

Formålet med dette initiativ er at forbedre afrapporteringen af studier om diagnostiske undersøgelers træfsikkerhed. Udtrykket diagnostisk træfsikkerhed (*diagnostic accuracy*) refererer til en tests egnethed til at identificere en given klinisk tilstand. En test er en procedure, som udføres for at få yderligere oplysninger om en patients helbredstilstand. Den tilsigtede forbedring af rapporterne skal føre til fyldestgørende beskrivelser, så læseren får præsenteret et tilstrækkeligt grundlag til at vurdere bias og andre mulige kilder til begrænsning af resultaternes almene gyldighed. STARD-initiativets vigtigste produkt er en checkliste (**Figur 1**), hvis tilblivelse og funktion omtales nærmere i det følgende.

Studier af diagnostiske tests træfsikkerhed går ud på at sammenligne resultaterne fra en eller flere test med resultater fra en referencestandard (guldstandard). De relevante undersøgelser udføres som reaktion på et genkendeligt klinisk problem, f.eks. mistanke om dyb venetrombose, som kræver en klinisk intervention. Testen kan være et spørgeskema, en blodprøve, en funktionsmåling eller billeddiagnostik, som udføres for at påvise eller udelukke den pågældende tilstand. Referencestandard kan være en undersøgelsestype, en gruppe af undersøgelser eller resultater fra opfølgende kliniske observationer. Studieobjektets (indekstestens) træfsikkerhed udtrykkes som graden af overensstemmelse mellem oplysninger fra indekstest og referencestandard. Træfsikkerhed kan beskrives kvantitativt ved hjælp af parametre som sensitivitet, specificitet, *likelihood*- og odds-ratioer, positive og negative prædiktive værdier, areal under ROC-kurver m.m.

Hvad er det for metodologiske mangler, som kritikerne har peget på, og hvad kan der gøres for at afhjælpe dem? Det har vist sig, at publikationer fra selv ansete videnskabelige tidsskrifter kan mangle elementære oplysninger om studiedesign, patientpopulation, metodevalg og databehandling [6]. En metodologisk svag publikation kan vildlede den ukritiske læser og være ubrugelig som kildemateriale, specielt til metaanaly-

## VIDENSKAB OG PRAKSIS | STATUSARTIKEL

Afsnit og emne	Punkt		Se side
Titel/resumé/nøgleord	1	Klassificer artiklen som en undersøgelse af diagnostisk træfsikkerhed (anbefalet MeSH overskrift: »sensitivity and specificity«)	
Indledning	2	Anfør studiets hypotese/formål, f.eks. vurdering af diagnostisk træfsikkerhed, sammenligning af træfsikkerhed mellem flere test eller patientgrupper	
Metoder Deltagere	3	<i>Beskriv:</i> Studiepopulationen: inklusions- og eksklusionskriterier, studiets placering og rammer	
	4	Rekruttering: var rekruttering baseret på debutsymptomer, tidligere undersøgelsesresultater eller at deltagerne var undersøgt med indekstest/referencestandard?	
	5	Udvælgelse: var studiepopulationen en konsekutiv serie af deltagere som defineret i punkt 3 og 4? Hvis ikke beskrives hvorledes udvælgelse fandt sted	
	6	Dataindsamling: blev dataindsamling planlagt før indekstest og referencestandard blev udført (prospektivt studie) eller efter (retrospektivt studie)?	
Testmetoder	7	Referencestandard med begrundelse/forklaring	
	8	Tekniske specifikationer vedr. materiale og metoder, anfør hvordan og hvornår målingerne blev udført, angiv referencer til indekstest og referencestandard	
	9	Definitioner af og rationale for måleenheder, cut off-værdier og klassifikationer for resultater fra indekstest og referencestandard	
	10	Antal, uddannelse og kompetence for de personer, som udførte og aflæste indekstest og referencestandard	
	11	Oplys om de personer, som aflæste indekstest/referencestandard, var hindrede i at have adgang til andre testresultater (blinding) og beskriv, hvilke kliniske informationer der var til deres rådighed	
Statistiske metoder	12	Metoder til beregning og sammenligning af diagnostisk træfsikkerhed og til angivelse af usikkerhed (f.eks. 95% konfidensinterval)	
	13	Metoder til beregning af testreproducerbarhed, hvis anvendt	
Resultater Deltagere	14	<i>Beskriv:</i> Hvornår studiet blev udført, inklusive datoer for rekrutteringens begyndelse og afslutning	
	15	Klinisk og demografisk karakteristik af studiepopulationen (f.eks. alder, køn, spektrum af debutsymptomer, komorbiditet, løbende behandling, rekrutteringscentre)	
	16	Anfør antal deltagere som imødekom inklusionskriterierne, som fik/ikke fik udført indekstest/referencestandard; beskriv årsager til manglende test (et flowdiagram kan stærkt anbefales)	
Testresultater	17	Anfør tidsinterval mellem indekstest og referencestandard og den i tidsrummet givne behandling	
	18	Fordeling af sygdomsstadier (efter definerede kriterier) blandt målgruppen; andre diagnoser blandt deltagere i referencegrupper	
	19	Krydstabulering af resultat af indekstest (inklusive ubestemmelige eller manglende data) versus resultat af referencestandard; ved kontinuerte data vises fordelingen af indekstestresultater versus referencestandardresultater	
	20	Utilsigtede hændelser i forbindelse med udførelse af indekstest/referencestandard	
Estimer	21	Estimation af diagnostisk træfsikkerhed, mål for statistisk usikkerhed (f.eks. 95% konfidensinterval)	
	22	Håndtering af ubestemmelige eller manglende indekstestresultater og outliers	
	23	Estimer af variabilitet vedr. diagnostisk træfsikkerhed mellem centre, testudførere og undergrupper af patienter, hvis udført	
	24	Estimer af testreproducerbarhed, hvis udført	
Diskussion	25	Diskuter den kliniske anvendelighed af studiets resultater	

Figur 1. STARD-checklisten til brug for forbedret rapportering af studier om diagnostiske undersøgelser træfsikkerhed. (Forklaring og uddybning: se [4]).

ser og andre oversigter. *Lijmer et al* har ydermere påvist, at bestemte typer af metodologiske mangler er relateret til overvurderinger af den pågældende tests diagnostiske ydeevne [7]. Der er således et klart behov for forbedring, hvilket flere internationale professionelle grupper har erkendt.

Man ved fra CONSORT-initiativet, at en positiv effekt kan opnås ved anvendelse af en checkliste, som viser de vigtigste metodologiske elementer, man i en rapport bør forholde sig til [2]. Tidsskriftet *Clinical Chemistry* har siden januar 2001 i sin forfatterinstruks inddraget en checkliste, som specielt ta-

## VIDENSKAB OG PRAKSIS | STATUSARTIKEL

- Diagnostiske studier kritiseres for svag metodologi og manglende konsekvensvurdering.
- Forbedringer skal opnås ved at tydeliggøre metodekrav, styrke tidskrifternes *review*-processer og fremme *outcome*-studier.

ger sigte på at forbedre kvaliteten af manuskripter om den diagnostiske træfsikkerhed af laboratorieundersøgelser [8]. STARD-initiativet har ført til en mere bredt anvendelig og bedre understøttet checkliste, der er sikret hurtig udbredelse gennem deltagerkredsen, som foruden repræsentanter for epidemiologisk, biostatistisk og forskningsmæssig ekspertise også tæller chefredaktører fra indflydelsesrige tidsskrifter som JAMA, BMJ og The Lancet. Checklisten har fået sin nuværende form (Figur 1) efter gennemgang af et omfattende kildemateriale, fremlæggelse og drøftelse ved STARD-gruppemødet i Amsterdam i september 2000 og efterfølgende yderligere modifikationer på grundlag af bl.a. bidrag via projektets hjemmeside, som fortsat vil være åben for kritiske kommentarer og forbedringsforslag.

STARD-gruppens materiale omfatter foruden checklisten også et prototypisk flowdiagram, idet det kraftigt anbefales at tydeliggøre studiets design og strømmen af undersøgte patienter ved hjælp af en figur. I ledsagedokumentet [4] redegøres for rationale bag de enkelte punkter i checklisten og for den underbyggende evidens.

#### »The evidence base of clinical diagnosis«: manual for diagnostisk forskning

En oversigt over bogens indhold kan findes i en anmeldelse af bogen i Ugeskriftet [9]. Her skal et af bogens budskaber fremhæves: Set fra patientens synspunkt er det vigtigere at vide, om en undersøgelse kan føre til et bedre behandlingsforløb og resultat, end om den bidrager til en korrekt diagnose. Strategier og metoder, som kan bruges til at besvare spørgsmål om *outcome*, diskuteres i to kapitler om henholdsvis den randomiserede, kontrollerede undersøgelse, en strategi som kan og bør anvendes i diagnostisk forskning – og det såkaldte før-efter-studie, som sigter på at afklare virkninger af at indføre en ny test.

Som illustrative eksempler henviser vi til publikationer fra danske forskergrupper.

Lassen *et al* [10] gennemførte en randomiseret undersøgelse af i alt 500 patienter med dyspepsi for at sammenligne effekten af to undersøgelsesstrategier: tidlig endoskopi versus testning for *H. pylori* og efterfølgende eradikationsbehandling. Efter et års opfølgende undersøgelser kunne man konkludere, at de to undersøgelsesprocedurer var lige effektive og lige sikre for patienterne.

Hvas *et al* [11] stillede spørgsmålstegn ved den kliniske betydning af forhøjet methylmalonat i plasma. I alt 140 patienter med let til moderat forhøjet plasmamethylmalonat blev randomiseret til behandling med B<sub>12</sub>-vitamin eller placebo. Der

blev udført grundige kliniske undersøgelser ved interventions begyndelse og tre måneder senere. En sammenligning af symptomer fra relevante organsystemer viste ingen signifikante forskelle, hvorfor man konkluderede, at B<sub>12</sub>-vitamin-behandling af patienter med forhøjet plasmamethylmalonat har begrænset klinisk betydning, i det mindste på kort sigt.

#### Perspektiver

Det er tankevækkende, at der skulle gå ti år fra en undersøgelse (in casu methylmalonatundersøgelsen) blev stillet til rådighed og taget i brug, før et forskerhold påtog sig at søge oplysninger om dens behandlingsmæssige konsekvenser. Videnskabelige og teknologiske fremskridt skaber tilbud om nye diagnostiske metoder i et tempo, som langt overskrider den kliniske forskningskapacitet for kritisk evaluering af diagnostisk træfsikkerhed og *outcome*. Derfor er det vigtigt at forbedre og synliggøre metodologiske principper og metoder, som kan fremme kvaliteten af de studier, som faktisk udføres trods manglen på resurser til området. Mindst lige så vigtigt er det, at læger, der som professionel gruppe er hovedansvarlig for diagnostik, bliver forsynet med teoretisk baggrund og praktiske redskaber, som sætter dem i stand til kritisk at vælge og udnytte de diagnostiske undersøgelser, som bedst kan gavne den enkelte patient.

De to ovenfor omtalte initiativer er skridt i den rigtige retning, men virkelige fremskridt opnås kun ved at øge den professionelle indsats. Her har de videnskabelige selskaber, tidsskrifts-redaktørene og fondsbestyrelserne en vigtig opgave at udføre.

Korrespondance: Erik Magid, Klinisk Biokemisk Afdeling, Amager Hospital, Italiensvej 1, DK-2300 København S. E-mail: magid@dadlnet.dk

Antaget: 9. januar 2004  
Interessekonflikt: Ingen angivet

\*)Dansk Selskab for Klinisk Biokemis arbejdsgruppe til fremme af evidensbaseret anvendelse af klinisk biokemi.

#### Litteratur

1. Jørgensen PE, Madsen JS, Jørgensen LGM *et al*. Klaringsrapporterne 1996-2001. Ugeskr Læger 2004;166(i trykken).
2. Moher D, Jones A, Lepage L, for the CONSORT group. Use of the CONSORT statement and quality of reports of randomised trials. JAMA 2001;285: 1992-5.
3. Bossuyt PM, Reitsma JB, Bruns DE *et al*, for the STARD group. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ 2003;326:41-4.
4. Bossuyt PM, Reitsma JB, Bruns DE *et al*. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med 2003;138:W1-12.
5. Knottnerus JA, ed. The evidence base of clinical diagnosis. London: BMJ Books, 2002
6. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. JAMA 1995;274:645-51.
7. Lijmer J, Mol BW, Heisterkamp S *et al*. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061-6.
8. Bruns DE, Huth EJ, Magid E *et al*. Toward a checklist for reporting studies of diagnostic accuracy of medical tests. Clin Chem 2000;46:893-5.
9. Magid E, Knottnerus JA, ed. The evidence base of clinical diagnosis. Ugeskr Læger 2002;164:5044.
10. Lassen AT, Pedersen FM, Bytzer P *et al*. Helicobacter pylori test-and-eradicate versus prompt endoscopy for management of dyspeptic patients: a randomized trial. Lancet 2000;356:455-60.
11. Hvas AM, Ellegaard J, Nexø E. Vitamin B<sub>12</sub> treatment normalizes metabolic markers but has limited clinical effect: a randomized placebo-controlled study. Clin Chem 2001;47:1396-404.