

Göran Lindstedt, professor emeritus (klinisk kemi), Sahlgrenska akademien, Göteborgs universitet
(goran.lindstedt.gu@telia.com)

Mats Eliasson, docent, institutionen för folkhälsa och klinisk medicin, Umeå universitet, överläkare i endokrinologi, Sunderby sjukhus, Luleå, medicinsk redaktör, Läkartidningen (mats.eliasson@nll.se)

Nya riktlinjer ger bättre dokumentation av undersökningsmetoder och resultat

II Två artiklar i JAMA från mitten och slutet av 1990-talet redovisade betydande kvalitetsbrister i vetenskapliga artiklar inom medicinen. Reid och medarbetare granskade artiklar om undersökningsmetoder (»tests«) publicerade i fyra ledande medicinska tidskrifter under en tidrymd av 16 år. Av 1 302 studier fyllde 112 kriterierna för vidare granskning. Av dessa befanns i medeltal bara en av fyra uppfylla uppställda krav på riktighet (»accuracy«) som acceptabel redovisning av patientmaterial och av mätnoggrannhet [1]. Det rörde sig här främst om radiologiska och kliniskt fysiologiska undersökningar samt immunkemiska och andra biokemiska mätningar. I senare studier av »diagnostic tests« har en vidare definition använts (Fakta 1).

Lijmer och medarbetare visade i en liknande granskning av 184 artiklar (218 mätmetoder) att brister i studiedesign var vanligt förekommande, ledande till överdrivet positiv bedömning av de aktuella mätningarnas diagnostiska värde [2].

Resultaten från seriösa granskningar av publikationer, även i ledande tidskrifter, talar alltså för att forskning inom detta område bör genomföras och dokumenteras bättre än vad som skett. Aktuella erfarenheter av systematiska översikter av diagnostiska metoder från SBU:s ventrombosgrupp talar för avsevärda kvalitetsbrister, som ofta omöjliggör tillförlitliga slutsatser om metodernas prestanda, interna validitet (bias) och externa validitet (mått på diagnostisk riktighet; generaliserbarhet och lämplighet). Därmed blir det kritiskt sammanställda vetenskapliga underlaget oftast alltför bristfälligt när sjukvården ska ta ställning till införande eller utrangerande av något diagnostiskt test. Hälsoekonomiska analyser saknas dessutom, vilket ytterligare försvårar möjligheten att skapa ett underlag för prioritering.

I en aktuell rapport [3] har man granskat faktorer som påverkar utfall och bedömning av mätmetoders prestanda inom medicinsk diagnostik.

Strikta krav på struktur och systematisk redovisning

Den evidensbaserade medicinen upptäckte tidigt svårigheter med att klarlägga behandlingsmetoders effekter och effektivitet på grund av brister i den vetenskapliga rapporteringen. År 1996 enades företrädare för de stora medicinska tidskrifterna om strikta krav på redovisning av metoder och resultat i randomiserade kliniska prövningar, the CONSORT statement [6]. Resultatet har varit att redaktörer och läsare snabbt kan värdera studiers tillförlitlighet och huvudresultat. De goda erfarenheterna ledde till liknande överenskommelser om rapportering av metaanalyser av randomiserade studier [7]

Sammanfattat



Internationella överenskommelser reglerar rapporteringen av randomiserade kliniska prövningar, systematiska översikter och observationsstudier.

Granskningar av artiklar om undersökningsmetoder och deras resultat har påvisat fundamentala brister, även i artiklar publicerade i ledande medicinska vetenskapliga tidskrifter.

Det finns därför ett behov av åtgärder för att öka trovärdigheten vid rapportering av mätmetoders prestanda och validitet, liksom av medicinska undersökningar (test) i stort.

År 2003 publicerades STARD-dokumentet för redovisning av undersökningar (Standards for Reporting of Diagnostic Accuracy). Ett stort antal tidskrifter har anslutit sig till de här nämnda kraven på manuskript.

Detaljerade riktlinjer för resultatdokumentation finns endast för ett mindre antal metoder. Det är ett krav på bl a laboratoriemedicinen att sådana riktlinjer sammanställs och blir allmänt tillgängliga.

Rapporter om diagnostiska metoder måste förbättra redovisningar av studiepopulation och studiedesign.

Vinsterna blir bättre kvalitet på vetenskapliga publikationer, större insyn, ökade möjligheter att klarlägga vetenskaplig oredlighet samt förbättrad forskarutbildning.

och för metaanalyser av observationsstudier [8]. Liknande riktlinjer finns även för hälsoekonomiska studier, vilka Läkartidningen nu anslutit sig till (www.lakartidningen.se).

Mätmetoder och trovärdighet

En ledande företrädare för skandinavisk laboratoriemedicin, Willie Gerhardt i Helsingborg, skrev för 20 år sedan en översikt över grafiska och tabellariska metoder för datasamman-

ställning [9]. Denna översikt kan ses som ett led i en dansk tradition inom det hälsovetenskapliga området, där Wulffs klassiska bok *Rational diagnosis and treatment* [10] lade grunden till det som utvecklades vidare till klinisk epidemiologi vid McMaster [11] och på 1990-talet mynnade ut i EBM-konceptet. En monografi på området skrevs av Doug Altman, som sedermera blev den ledande statistikern inom Cochrane-samarbetet [12]. Aktuella översikter ges i flera dokument [13-15]. Det är emellertid viktigt att skapa trovärdighet när det gäller de enskilda mätmetodernas prestanda och validitet.

STARD-initiativet

För att klargöra mätningars validitet och relevans för den aktuella frågeställningen, baser för valda beslutsgränser samt relevans av valda studiepopulationer har enskilda tidskrifter utvecklat checklistor för författare och referenter för skrivning respektive bedömning av manuskript insända för publicering. En sådan lista, tillämpbar på mätningar inom laboratoriemedicinens alla fält, publicerades år 2000 [16]. Otvivelaktigt ledde denna till förbättringar [17, 18].

En internationell samarbetsgrupp, STARD (Standards for Reporting of Diagnostic Accuracy), utgick från ett stort antal (33) sådana checklistor, från vilka 75 punkter valdes ut för vidare bearbetning. I januari-februari 2003 publicerades det så kallade STARD-dokumentet tillsammans med kommentarer i en lång rad ledande tidskrifter inom laboratoriemedicin och andra medicinska discipliner [4, 19-21]. Sedan dess har alltför många tidskrifter anslutit sig till dessa rekommendationer. Bland annat har man framhållit »to enable readers of diagnostic research reports to evaluate whether methodological key issues were addressed, authors are advised to follow the STARD guidelines« [22]. Det har också framhållits att STARD-dokumentets checklista (Fakta 2) och flödesdiagram för beskrivning av studiedesign är väl lämpade även för andra undersökningar än kliniskt kemiska mätningar (Fakta 1) [23]. Den tidskrift som i skrivande stund senast anslutit sig är *Family Practice* [24]. Whiting och medarbetare [5] redovisar en checklista för kvalitetsbedömning av studier av diagnostisk riktighet som ett komplement till STARD-dokumentet.

Kritik har dock riktats mot STARD-dokumentet för att det i nuvarande form belyser mätosäkerhet ofullständigt [25].

Ramverk för resultatredovisning

STARD-dokumentet får ses som en stomme för redovisning av medicinska undersökningsresultat. De 25 punkterna i checklistan tar bl a upp studiepopulation och patientrekrytering, datainsamling, mätmetoder, statistiska metoder, resultatredovisning och resultatens kliniska användbarhet.

Redovisningen av undersökningsmetoder måste emellertid kompletteras med information som är specifik för det aktuella området. Här är litteraturen dock ofullständig. Feinstein efterlyste exempelvis år 2002 definition och dokumentation av omvårdnads-kvalitet [26, 27].

Det är väl förståeligt att intresset för att dokumentera kvaliteten är störst för ny metodik. Behovet av standarder för redovisning av så kallade molekylärgenetiska undersökningar framhölls av Feinsteins grupp 1999 [28]. År 2001 publicerades riktlinjer för beskrivning av mätningar av genuttryck med så kallad microarray-teknik (mikromatrisundersökning) [29]. Flera tidskrifter publicerade gruppens öppna brev i frågan [30]. Riktlinjerna antogs omgående av ledande tidskrifter inom fältet [31]. Nyligen publicerades riktlinjer för datapresentation från tvådimensionell elektrofores för så kallade proteomikundersökningar [32].

Immunkemiska mätningar och bristen på riktlinjer

Det finns ännu inga allmänt accepterade riktlinjer för beskrivning av immunkemiska mätningar, trots att denna grupp av

|| Fakta 1

Definition av »test« (undersökningsmetod) i artiklar om diagnostisk riktighet [4, 5]

Undersökning som syftar till att samla information om en individs hälsotillstånd, t ex anamnesupptagande, fysikalisk undersökning, frågeformulär, laboratorieundersökning inklusive funktionsundersökning, radiologisk och nuklearmedicinsk undersökning (imaging test, avbildningsundersökning), kirurgisk exploration och patologisk-anatomisk undersökning.

|| Fakta 2

Översikt över STARD-listan för redovisning av diagnostisk riktighet i 25 delavsnitt [4]

Titel/Abstrakt/Nyckelord.

Introduktion.

Metoder: Beskrivning av deltagare (4 delavsnitt), undersökningsmetoder (5 delavsnitt), statistiska metoder.

Resultat: Beskrivning av deltagare (ytterligare 3 delavsnitt), undersökningsresultat (3 delavsnitt), datasammanställning och bedömningar (4 delavsnitt).

Diskussion.

metoder har en vid användning inom medicinsk forskning och sjukvård. Ännu i början av 1970-talet beskrevs deras utförande som »more art than science«, vilket antyder de många faktorer som kan påverka resultat från sådana mätningar. Publikationer baserade på immunkemiska mätningar är legio. Det gäller även publikationer som saknar till och med basal metodologisk information, t ex om analytens sammansättning (dvs vad som mäts under de aktuella betingelserna), metodens kalibrering, mätvärdens riktighet och reproducerbarhet och fördelningen av mätvärdena hos sjuka och friska och hos patienter med samma symtomatologi men med annan diagnos. Endast sällan ser man att författarna klarlagt i vilken utsträckning som mätvärdena påverkats av analytisk interferens. Samma gäller analys av extremvärden (outliers). I avsaknad av sådan information torde det vara omöjligt att reproducera resultaten för andra forskare.

Exempel på faran av att använda immunkemiska metoder som inte validerats adekvat för klinisk användning är rapporterna från mätning av det aminoterminala fragmentet av proBNP, ofta benämnt N-terminalt proBNP, med en kommersiellt tillgänglig metod baserad på ett antiserum riktat mot en central region i molekylen. Vad som egentligen mättes i patientprov med den aktuella mätmetoden är obekant – men inte var det den uppgivna substansen [33].

Ett något mindre drastiskt exempel är mätningen av prolaktin i sjukvård och humanmedicinsk forskning. Ingen seriös tidskrift torde väl nu acceptera ett manuskript med redovisning av prolaktinmätningar, där författarna inte klargjort om de tagit hänsyn till antikroppar i prov, riktade mot prolaktin och ledande till så kallad makroprolaktinemi [34].

Det finns också andra exempel på polypeptider som cirkulerar i blodbanan i såväl fri som komplexbunden form, vilket

kan leda till problem vid tolkningen av mätvärdena. Ett snabbt expanderande sådant område, där immunkemisk mätning intar en central roll, är cytokiner. Kraven är här stora på adekvat metodinformation [35].

För information om beskrivning av immunkemiska mätningar hänvisas man alltså till monografier i ämnet och till de relativt få exempel på kvalitets-specifikationer som formulerats för enskilda grupper av substanser, exempelvis troponiner [36]. Det är också viktigt att författarna har granskat den/de aktuella substansens/-ernas patofysiologi för att klargöra vilka biologiska faktorer som kan tänkas påverka mätvärdena. Ofta saknas exempelvis information om njurfunktionen – för vissa komponenter spelar njurarna en viktig roll vad gäller elimineringen från blodbanan.

Studiepopulation och studiedesign

För rätt tolkning av mätvärdena krävs att studiepopulationen/-erna är adekvat/a för den studerade frågeställningen. För att kunna bedöma detta måste man definiera och dokumentera den/dem väl. Som nämns ovan stod det klart i mitten–slutet av 1990-talet att betydande brister förelåg härvidlag i den vetenskapliga litteraturen. STARD-dokumentet ägnar därför inte mindre än sju av 25 punkter åt dessa problem, från inklusions- och exklusionskriterier och sjukdomsspektrum till flödesschema för beskrivning av antal patienter som genomgått de skilda undersökningarna. Aktuella översikter beskriver orsaker till feltolkningar som följd av inadekvat studiedesign [3, 37]. I redovisningen av studiepopulationen bör också ingå redovisning av icke-tolkningsbara undersökningsresultat, som t ex vid ultraljudsundersökning vid samtidig förekomst av benödem eller andelen ej utvärderbara pulmonalisangiografier vid misstänkt lungemboli [14].

Preanalytiska och postanalytiska orsaker till fel

STARD-dokumentet belyser inte tillräckligt väl betydelsen av noggrann dokumentation av även preanalytiska förhållanden för att man ska kunna undvika systematiska fel (bias) beroende på instabilitet i de uppmätta substanserna, interferens vid mätningen, etc. Även postanalytiska orsaker till fel bör beaktas [38]. Håri ingår även analys av extremvärden, något som alltför sällan redovisas.

Bas för evidensbaserad laboratoriemedicin

Arbetet med riktlinjer för förbättrad dokumentation av vetenskapliga studier ingår i ansträngningarna att skapa en bas för evidensbaserad laboratoriemedicin [39, 40].

Slutsatser

Förbättrad redovisning av planering och genomförande av vetenskapliga studier inom medicinen, liksom av tolkning av resultaten och deras kliniska implikationer, kan leda till att

- kvaliteten och validiteten i vetenskapliga manuskript höjs, och bättre beslutsunderlag skapas för den kliniska verksamheten, vilket möjliggör kostnadseffektivitetsanalyser och prioriteringsdiskussioner
- tidskrifternas och referenternas bedömning av manuskript insända för publicering underlättas
- andra forskargrupper inom fältet får möjlighet att studera samma och närliggande problem om resultaten ifrågasätts
- forskningsetikkommittéernas arbete underlättas och effektiviseras
- anslagsbeviljande organisationers arbete för rättvis bedömning underlättas och effektiviseras
- möjligheterna ökar vad gäller att klarlägga vetenskaplig oredlighet och att frigöra sig från oberättigade anklagelser av detta slag (Fakta 3)
- forskarutbildningen förbättras.

II Fakta 3

Forskningsfusk

Vetenskaplig oredlighet (forskningsfusk eller »scientific fraud») är fortfarande ett aktuellt problem inom medicinsk och annan forskning [42-50]. Det finns betydande skillnader i uppfattningen om vilka insatser som bör göras för att upptäcka, åtgärda och förebygga forskningsfusk.

Adekvat dokumentation av forskningsresultat är emellertid en förutsättning för att såväl problematiken ska kunna klarläggas som oberättigade anklagelser om forskningsfusk ska kunna avvisas.

Vetenskapliga tidskrifter kan göra viktiga insatser

Rolf Zetterström och kolleger inom kretsen av tidskriftsredaktörer har insiktsfullt och tankeväckande beskrivit forskningsfusk med tonvikt lagd vid biokemiska och metabola undersökningar [44, 47].

Tesen är att de vetenskapliga tidskrifterna bör engageras i kampen mot forskningsfusk. Detta är desto mer betydelsefullt, eftersom universiteten alltför sällan tar sitt ansvar i dessa frågor – en uppfattning som finner stöd även i senare litteratur [51-54]. Bland annat framhålls: »Forskningsprotokoll måste också hållas tillgängliga för de vetenskapliga redaktionerna för den händelse de skulle behöva granskas i samband med bedömning av artiklar» [44].

Men för att referentgranskningssystemet ska kunna fungera krävs adekvat dokumentation av det vetenskapliga underlaget. Detta material skulle då vara tillgängligt för granskning av vederbörande myndighet, med anlitande av expertis inom området, i de fall utredning sker av misstänkt forskningsfusk.

Brister i vetenskapliga rapporter

Är då de vetenskapliga tidskrifterna garanter för god kvalitet av redovisningar av forskningsresultat? Framför allt artiklarna av Reid och medarbetare [1] och Lijmer och medarbetare [2] talar mot att så skulle vara fallet.

För att ge granskare av insända manuskript, liksom läsarna av publicerade artiklar, ökade möjligheter att klarlägga forskningsfusk bör tidskrifternas redaktioner kräva att författarna uppfyller de krav som uppställs i the CONSORT statement [6, www.consort-statement.org], COPE-riktlinjerna för publiceringsetik [45, www.publicationethics.org.uk], STARD (se text) och liknande dokument.

STARD-dokumentet och efterföljare är viktiga steg på vägen i denna riktning. En översyn av aktuella metoder inom sjukvården i denna anda ger också förutsättningar för bättre sjukvård. Som exempel kan nämnas metodiken för blodtrycksmätning i primärvård [41].

*

Potentiella bindningar eller jävsförhållanden: Inga uppgivna.

Referenser

1. 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.
2. Lijmer JG, Mol BW, Heisterkamp S, Bossel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;282:1061-6.
3. 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.
4. 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. Standards for Reporting of Diagnostic Accuracy. *Clin Chem* 2003;49:1-6.
5. 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. (<http://www.biomedcentral.com/1471-2288/3/25>)

6. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA* 1996;276:637-9. (se även www.consort-statement.org)
8. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. *JAMA* 2000;283:2008-12.
9. Gerhardt W, Keller H. Evaluation of test data from clinical studies. I. Terminology, graphic interpretation, diagnostic strategies, and selection of sample groups. II. Critical review of the concepts of efficiency, receiver operated characteristics (ROC), and likelihood ratios. *Scand J Clin Lab Invest Suppl* 1986;181:1-74.
10. Wulff HR, Gotzsche PC. Rational diagnosis and treatment. Evidence-based clinical decision-making. 3rd ed. Oxford: Blackwell; 2000.
13. Sasse EA. Objective evaluation of data in screening for disease. *Clin Chim Acta* 2002;315:17-30.
14. Statens beredning för medicinsk utvärdering. 3.1 Utvärdering av diagnostiska metoder. I: Blodpropp – förebyggande, diagnostik och behandling av venös tromboembolism: en systematisk kunskapsammanställning. Vol 2. Stockholm: SBU; 2002. p. 13-27.
17. Bossuyt PM. The quality of reporting in diagnostic test research: getting better, still not optimal. *Clin Chem* 2004;50:465-6.
19. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin Chem* 2003;49:7-18.
20. Bruns DE. The STARD initiative and the reporting of studies of diagnostic accuracy. *Clin Chem* 2003;49:19-20.
36. Panteghini M, Gerhardt W, Apple FS, Dati F, Ravkilde J, Wu AH. Quality specifications for cardiac troponin assays. International Federation of Clinical Chemistry and Laboratory Medicine (IFCC). IFCC Scientific Division Committee on Standardization of Markers of Cardiac Damage. *Clin Chem Lab Med* 2001;39:175-9.
40. Oosterhuis WP, Bruns DE, Watine J, Sandberg S, Horvath AR. Evidence-based guidelines in laboratory medicine: principles and methods. *Clin Chem* 2004;50:806-18.
44. Zetterström R. Glöm inte tidskrifterna i kampen mot forskningsfusk. *Läkartidningen* 1998;95:4864-7.
45. Committee on Publication Ethics (COPE). Guidelines on good publication practice. <http://www.publicationethics.org.uk>
47. Zetterström R, Mitelman F, Rydholm A. Vetenskapligt fusk – tidskrifternas dilemma. *Läkartidningen* 2004;101:2743-4.
50. Wilmshurst P. Obstacles to honesty in medical research. *HealthWatch Newsletter*, 2004(52)(January) [accessed 5 February 2005]. <http://www.healthwatch-uk.org/hw52.html#wilmshurst>



I Läkartidningens elektroniska arkiv
<http://tarkiv.lakartidningen.se>
är artikeln kompletterad med fullständig referenslista



=artikeln är referentgranskad

SUMMARY

Surveys of published studies on diagnostic accuracy and the performance of diagnostic tests, widely defined as any method for obtaining additional information on the health status of an individual, have revealed poor methodological quality. This holds true also for scientific journals considered to be leading in their fields. The STARD document (Standards for Reporting of Diagnostic Accuracy), which was first published in 2003, provides guidelines for the design, performance and report of such studies. Adherence to these guidelines may improve the quality of scientific publications, increased opportunities for readers to get an insight into the experimental details, improved education in medical research and increased opportunities for the disclosure of scientific fraud.

Göran Lindstedt, Mats Eliasson

Correspondence: Göran Lindstedt, Sahlgrenska akademien, SE-413 45 Göteborg Sweden (goran.lindstedt.gu@telia.com)