Utvärdering av metoder i hälso- och sjukvården

En handbok

SBU — STATENS BEREDNING FÖR MEDICINSK UTVÄRDERING

www.sbu.se/metodbok

Diarienummer STY2013/5 • Version 2013-05-16

SBU – Statens beredning för medicinsk utvärdering, Stockholm

Redaktör Måns Rosén
Grafisk produktion Elin Rye-Danjelsen

Vänligen citera denna publikation så här:
SBU. Utvärdering av metoder i hälso- och sjukvården: En handbok. Version 2013-05-16 Stockholm: Statens beredning för medicinsk utvärdering (SBU). Hämtad från www.sbu.se/metodbok den [datum].

Innehåll

Bilagor

Förord

Detta är första upplagan av en metodbok kring SBU:s huvuduppgift, dvs systematisk granskning av metoder i hälso- och sjukvården. Syftet är främst att vara en vägledning för experter i SBU:s projekt och för medarbetare om hur vi ska bedriva granskningsarbetet på ett systematiskt, enhetligt och öppet sätt. Eftersom intresset för utvärdering av metoder i hälso- och sjukvården (internationellt benämnt ”health technology assessment”) har ökat kraftigt såväl nationellt som regionalt och lokalt, ser vi även att en bredare målgrupp kan vara intresserad av att veta hur sådant arbete bedrivs.

Arbetet med att utveckla denna metodbok har drivits internt, med externa experter och i diskussioner med SBU:s råd och SBU:s nämnd. Den metodik som redovisas i denna rapport ska tillämpas i SBU:s samtliga projekt. Metodboken ska vara ett levande dokument som med jämna mellanrum kan revideras efter att nya erfarenheter har dragits i projektarbetet.

Stockholm i februari 2013

Måns Rosén
DIREKTÖR SBU

1. Utvärdering av metoder i hälso- och sjukvården – inledning

VERSION 2010:1

Evidensbaserad vård

Hälso- och sjukvården har haft en snabb utveckling som alltmer baseras på vetenskapliga rön. Kraven på att tillämpa behandlingar som har vetenskapligt stöd för effekt har ökat. Begreppet *EBM* (”evidensbaserad medicin” eller ”evidensbaserad vård”) kan ses som ett uttryck för detta. EBM är ett förhållningssätt där man hela tiden kritiskt bedömer om vården vilar på bästa tillgängliga vetenskapliga grund [1]. Det medför i sin tur att vårdgivaren måste kunna bedöma vad som är bästa tillgängliga metod.

Det blir allt svårare att hinna hålla sig uppdaterad inom sitt verksamhetsområde. Mängden artiklar som publiceras per år ökar kontinuerligt. Beräkningar visar att drygt 1,4 miljoner medicinska artiklar publiceras årligen – och av dem uppskattas cirka 10–15 procent ha ett praktiskt och bestående värde för patienterna.

Systematisk översikt

Ett sätt att få kunskapen sammanfattad är att läsa en *översikt* (”review”). Svagheten med icke-systematiska översikter är att de ofta bygger på studier som författaren känner till. Dessutom finns det risk för att författaren väljer ut enbart de studier som stödjer författarens egna åsikter. Översikten kan därför komma att ge en skev bild av de verkliga förhållandena.

En *systematisk översikt* (”systematic review”) ska uppfylla höga krav på tillförlitlighet. En bra systematisk översikt följer vissa principer som ska minimera riskerna för att slumpen eller godtycklighet påverkar slutsatserna. Hit hör:

- En preciserad fråga/problem.
- Reproducerbarhet: redovisning av urvalskriterier (inklusions- och exklusionskriterier) för att sålla fram den relevanta litteraturen samt strategier för sökning och kvalitetsgranskning.
- Systematisk sökning efter all relevant litteratur för den fråga eller problem som behandlas.
- Kvalitetsgranskning av samtliga studier som uppfyller urvalskriterierna.
- Extraktion av data och tabellering från de studier som har kvalitetsgranskats.
- Sammanvägning av resultaten i t ex en metaanalys.
- En bedömning av hur välgrundade resultaten är (evidensgradering).

En välgjord systematisk översikt ger läsaren möjlighet att bedöma trovärdigheten i slutsatserna och att kontrollera om någon viktig litteratur inte kommit med i bedömningen.

Utvärdering av metoder i hälso- och sjukvården

1

Utvärdering av metoder i hälso- och sjukvården (”health technology assessment”) står för en systematisk utvärdering av det vetenskapliga underlaget för metodernas effekter, risker och kostnader, [2]. Det gäller för alla metoder som används vare sig det gäller prevention, diagnostik, behandling eller omvårdnad. Den systematiska översikten av effekter, risker och kostnader ska enligt SBU:s uppdrag kompletteras genom att även väga in etiska och sociala aspekter. Utvärderingen har en bredare ansats och kommer därmed att ta mer hänsyn till de nationella/lokala förhållandena än en systematisk översikt.

Faktaruta 1.1 SBU – en av världens äldsta organisationer för utvärdering av hälso- och sjukvårdens metoder.

SBU har regeringens uppdrag att göra fullständiga utvärderingar av metoder som används inom hälso- och sjukvården. Resultaten av utvärderingarna ska vara vägledande för såväl de praktiska utövarna som för politisk och administrativ ledning på olika nivåer. I uppdraget ingår också att sprida resultaten från dessa utvärderingar till hälso- och sjukvården i Sverige och följa upp effekterna av dessa insatser. SBU är sannolikt numera världens äldsta existerande nationella organisation för utvärdering av medicinska metoder. SBU startade sin verksamhet 1987.

Val av ämnen för utvärdering

SBU får in förslag till utvärderingar från många håll. De kan komma från t ex hälso- och sjukvårdspersonal, specialistföreningar, landstingsledningar och andra myndigheter inom hälso- och sjukvårdsområdet. En del utvärderingar blir underlag för Socialstyrelsens nationella riktlinjer eller Tandvårds- och läkemedelsförmånsverkets beslut.

De projektförslag som kommer in rangordnas med hjälp av ett antal kriterier. Ju fler kriterier som uppfylls, desto mer angelägen är frågan. Kriterierna är:

- stor betydelse för liv och hälsa
- vanligt hälsoproblem – berör många
- stor variation i praxis
- ofullständig kunskap om hur starkt det vetenskapliga underlaget är
- stora ekonomiska konsekvenser

- viktig etisk fråga
- stor betydelse för organisation eller personal
- kontroversiell eller uppmärksammad fråga.

SBU:s råd och Alertrådet granskar den vetenskapliga kvaliteten i SBU:s rapporter. SBU:s nämnd beslutar vilka projekt som ska genomföras och ska stå bakom de slutsatser som dras i rapporterna. Nämnden består av företrädare för centrala organisationer inom hälso- och sjukvården i Sverige. Nämndens sammansättning ska garantera att projekten har en bred förankring, anses betydelsefulla och att slutsatserna är väl förankrade i svensk hälso- och sjukvård.

Faktaruta 1.2 Ansvarsfördelning mellan hälso- och sjukvårdsmyndigheter i Sverige.

Det finns ett väl fungerande samarbete mellan hälso- och sjukvårdsmyndigheterna i Sverige med avgränsade ansvarsuppgifter som något förenklat kan beskrivas enligt nedan:

Myndighet	Huvuduppgifter
Läkemedelsverket	Beslutar om godkännande av läkemedel. Fokus på effekt och säkerhet
Tandvårds- och läkemedelsförmånsverket (TLV)	Beslutar om subventionering av läkemedel och tandvård. Fokus på kostnadseffektivitet
SBU	Ansvar för systematiska kunskapsöversikter och utvärdering av metoder i hälso- och sjukvården
Socialstyrelsen	Ansvar för nationella riktlinjer, föreskrifter, register m m

Ämnesexperterna har en central roll i arbetet

Utvärderingen görs av experter inom ett ämnesområde med stöd från SBU:s kansli. Detta skiljer SBU från många andra organisationer som tar fram systematiska översikter och medicinska utvärderingar. Ofta sammanställs deras rapporter av experter på själva granskningsmetodiken och de har liten möjlighet att bedöma vilken klinisk relevans en metod har. Experterna i SBU:s projekt säkrar att utvärderingen grundas på djup förståelse för ämnesområdet.

Det är viktigt att projektgruppen är allsidigt sammansatt. Frågan som ska utredas berör ju oftast flera yrkeskategorier. Projektgruppen bör också ha en könsmässig och geografisk spridning.

Arbetsprocessen

Större projekt som omfattar ett helt sjukdomsområde tar några år att genomföra. Det finns en del kontrollstationer, där arbetet presenteras för de vetenskapliga råden för diskussion. När manus är klart väntar en omfattande granskning. Utkastet granskas av en intern kvalitetsgrupp som i huvudsak bedömer den metodologiska kvaliteten. Det skickas också till flera externa granskare som i första hand bedömer om innehållet är relevant. Manus tas därefter upp i något av de vetenskapliga råden. När rapporten godkänts av dem går den till SBU:s nämnd för synpunkter och beslut om publicering. Denna process får nog sägas vara mer omfattande än den som föregår publicering i vetenskapliga tidskrifter.

Lästips och läsanvisningar

Det finns flera svenska [3–8] och internationella [9–13] publikationer som ger en grundläggande eller mer fördjupad beskrivning av evidensbaserad medicin/omvårdnad och utvärdering av metoder i hälso- och sjukvården (”health technology assessment”).

Metodboken följer de olika stegen i arbetsprocessen. Den kan läsas i en följd eller användas som uppslagsbok av experterna i olika skeden av ett projekt.

Metodboken inleds med en översikt över de olika stegen i en systematisk granskning och utvärdering (Kapitel 2). Därefter följer ett avsnitt om formulering av frågor och val av selektionskriterier (Kapitel 3) följt av litteratursökning och val av databaser (Kapitel 4). Bedömning av studiens relevans beskrivs i Kapitel 5. Kvalitetsgranskning av studier med olika studiedesign beskrivs i Kapitel 6 och 7. Kapitel 8 tar upp utvärdering av kvalitativa studier. Användning av metaanalyser finns i Kapitel 9. I Kapitel 10 redovisas hur den sammanfattande evidensgraderingen ska göras. Hälsoekonomi återfinns i Kapitel 11, och i Kapitel 12 tas etik och sociala aspekter upp.

Längst bak i metodboken finns bilagor med de olika mallarna som används vid granskning av studier. Grundläggande statistiska begrepp redovisas vidare i Bilaga 9.

Referenser

1. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence-based medicine: what it is and what it isn't. BMJ 1996;312:71-2.

2. Banta D, Jonsson E, editors. History of HTA. Int J Technol Assess Health Care 2009;25 suppl 1:1-289.

3. Brorsson B, Wall S. Värdering av medicinsk teknologi – problem och metoder. Stockholm: Medicinska forskningsrådet; 1985.

4. Nordenström J. Evidensbaserad medicin i Sherlock Holmes fotspår. 4:e upplagan. Karolinska University Press; 2007.

5. Furberg B, Furberg C. Allt är inte guld som glimmar. III Hur man värderar kliniska studier. Kungsbacka: Solutio; 2005.

6. Larsson A. Arbetsbok i evidensbaserad medicin. Södra Älvsborgs sjukhus 2006:2.2.

7. Willman A, Stoltz P, Bahtsevani C. Evidensbaserad omvårdnad. En bro mellan forskning och klinisk verksamhet. Studentlitteratur; 2006.

8. Levi R. Vettigare vård. Evidens och kritiskt tänkande i vården. Stockholm: Norstedts; 2009.

9. Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions. Version 5.0.0 (update February 2008), Cochrane collaboration 2008. Available from www.cochrane-handbook.org.

10. Fletcher RH, Fletcher SW. Clinical epidemiology. The essentials. 4th ed. Lippincott Williams & Wilkins: Baltimore; 2005.

11. Guyatt G, Rennie D, editors. User's guide to the medical literature. A manual for evidence-based clinical practise. JAMA & Archives Journal; 2002.

12. Egger M, Smith DG, Altman DG, editors. Systematic reviews in health care: meta-analysis in context. London: BMJ Books; 2001.

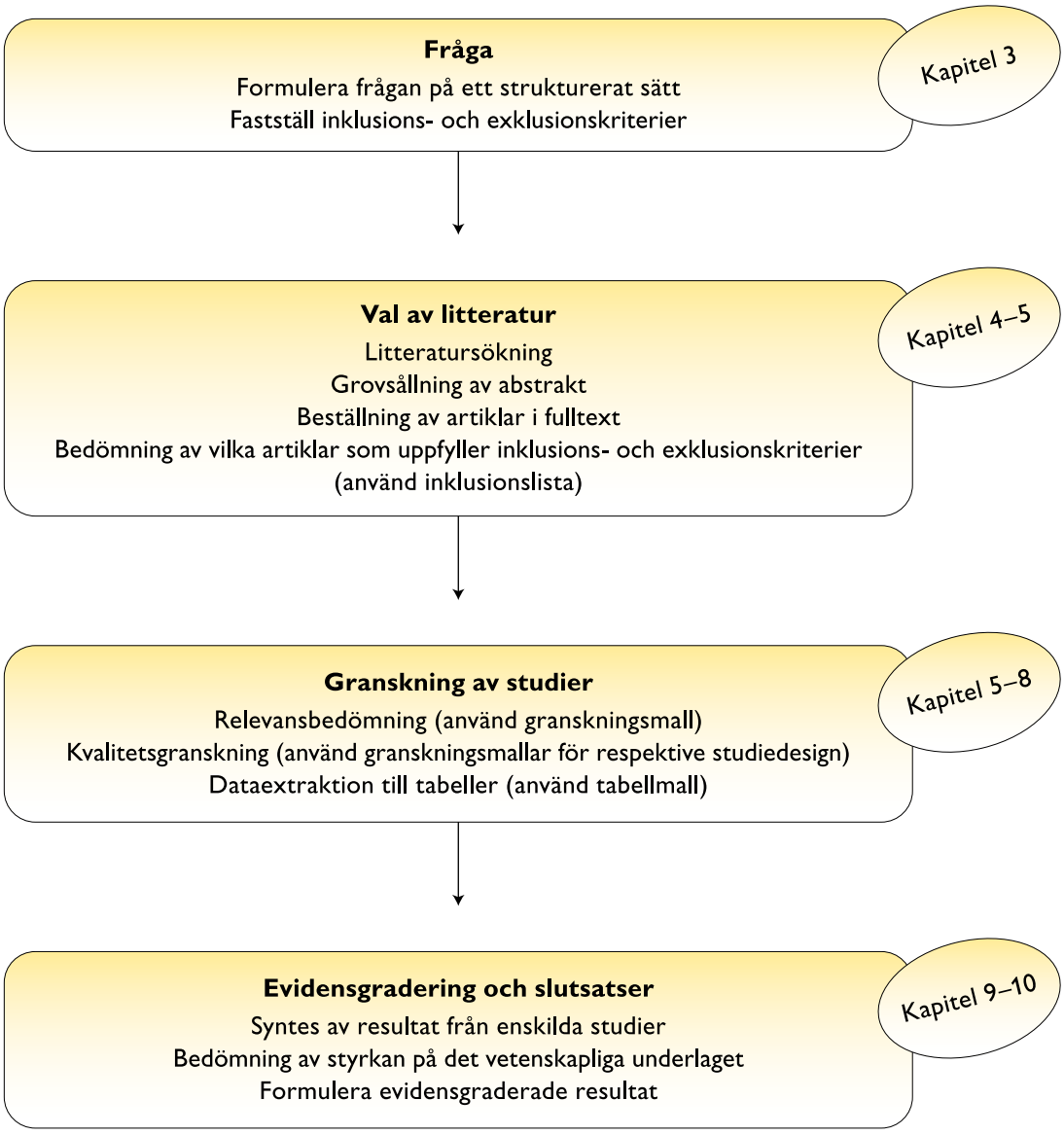
13. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology. A basic science for clinical medicine. 2nd ed. Little, Brown and company: Boston; 1991.

2. En översikt av stegen i en systematisk utvärdering

VERSION 2012:1

Inledning

Den metod för utvärdering som SBU tillämpar grundas på en systematisk granskning av den vetenskapliga litteraturen. Detta innebär att sökningen av relevant litteratur, urval och kvalitetsgranskning görs på ett systematiskt sätt. Det är viktigt att varje fas i processen är väl definierad och tydligt redovisad i rapporten (Figur 2.1). I detta avsnitt redovisar vi en sammanfattning av vad som ingår i de olika delarna av utvärderingen.



Figur 2.1 Process för systematisk utvärdering av vetenskapligt underlag.

Formulering av frågorna i projektet (Kapitel 3)

De frågor som projektet omfattar är till en början i regel formulerade på en generell nivå. En första uppgift för projektgruppen är därför att strukturera frågorna så att de kommer att kunna besvaras. Detta arbete är avgörande för vilka studier som kommer att fångas in i litteratursökningen och måste därför göras omsorgsfullt.

2

Projektgruppen ska ta ställning till vilka populationer som är intressanta för respektive fråga, vilka metoder som ska utvärderas inom projektets ram och vilka utfallsmått som ska studeras. I de allra flesta fall definieras även vilka kontrollmetoder som är relevanta. Frågan formuleras därefter enligt det så kallade PICO-formatet (”population, intervention, control, outcome”) för interventionsstudier och enligt PIRO-formatet (population, indextest, referenstest, ”outcome”) för studier om diagnostisk säkerhet.

Frågorna specificeras ytterligare med hjälp av inklusions- och exklusionskriterier.

Litteratursökning (Kapitel 4)

Litteratursökningen genomförs av en informationsspecialist i samråd med projektets experter och projektledare. Experternas roll är framför allt att bidra med relevanta artiklar till informationsspecialisten som analyserar abstrakt och indexering för att utveckla sökstrategin. Experterna ger också förslag på, för ämnesområdet, lämpliga termer till sökstrategin.

För att minimera risken för systematiska fel när det gäller litteratursökningen utförs sökningar i flera databaser samt att kompletterande kontroll av referenslistor görs. Målet med sökstrategierna är att om möjligt fånga alla relevanta studier samtidigt som antalet icke relevanta artiklar är så få som möjligt. I slutfasen av ett projekt görs en uppdaterande sökning för att inkludera artiklar som publicerats under projekttiden.

Sökstrategierna inklusive sökresultaten redovisas i rapporten.

Bedömning av en studies relevans (Kapitel 5)

Två personer (eller fler) granskar, oberoende av varandra, abstraktlistor från databassökningarna. Studier som bedöms vara relevanta utifrån rubrik och abstrakt beställs i fulltext. Det räcker att en av granskarna anser att artikeln bör läsas i fulltext för att den ska beställas.

Antalet beställda artiklar ska redovisas i rapporten.

De två granskarna bedömer därefter, oberoende av varandra, om de beställda artiklarna uppfyller inklusionskriterierna. Som stöd i arbetet används ett formulär för inklusion och exklusion. De artiklar som inte uppfyller kriterierna sorteras bort. I formuläret anges orsaken till exklusion. Granskarna jämför därefter sina inklusionslistor. Om listorna inte överensstämmer, diskuterar granskarna inbördes och beslutar huruvida artikeln ska inkluderas eller inte.

Rapporten ska innehålla en redovisning av antalet exkluderade artiklar och orsak till exklusion.

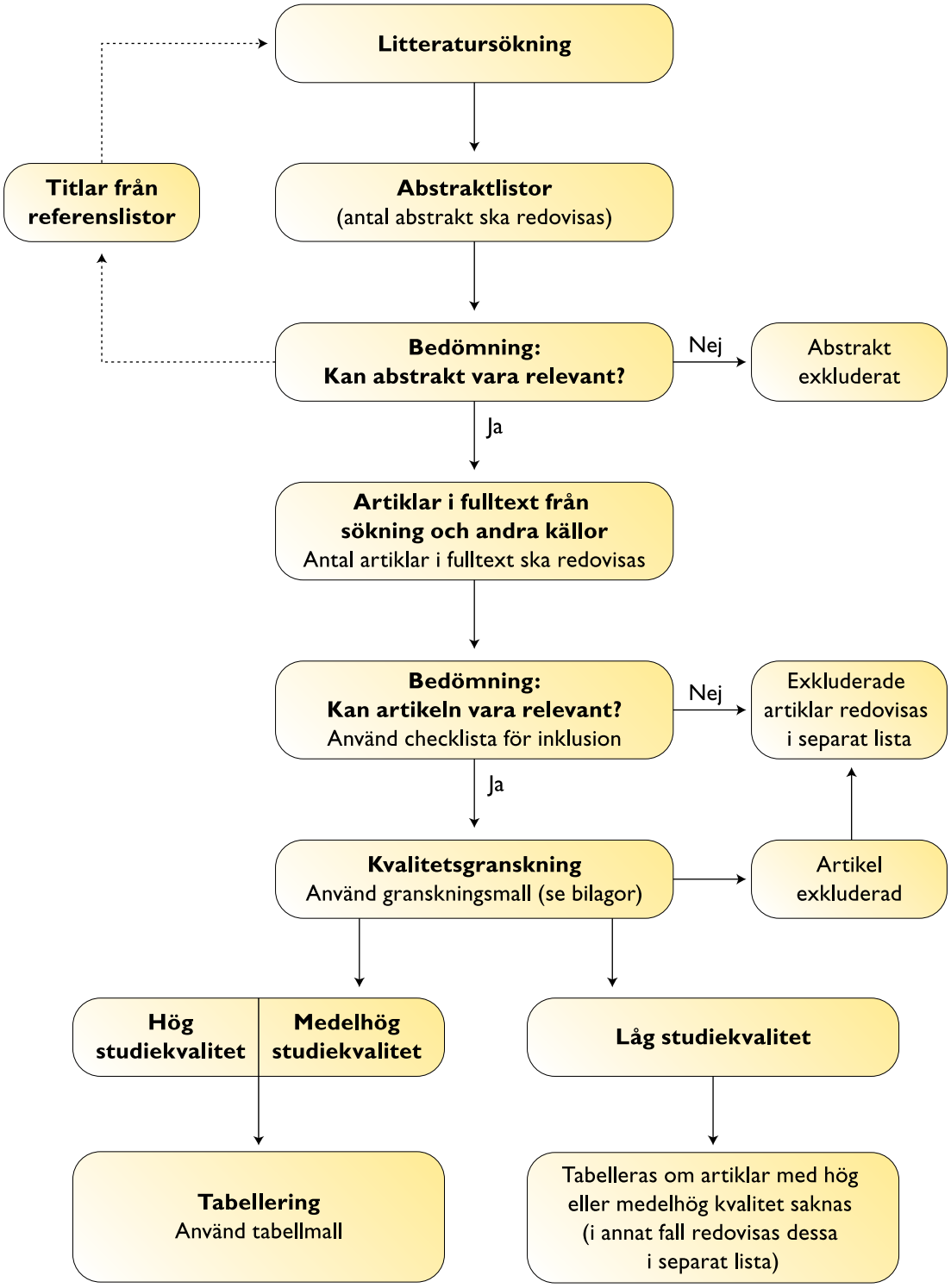
Urvalsprocessen från litteratursökning till tabellering kan sammanfattas enligt Figur 2.2.

Kvalitetsgranskning och dataextraktion (Kapitel 6–8)

I nästa steg bedömer granskarna, oberoende av varandra, kvaliteten på de preliminärt inkluderade studierna. Som stöd för arbetet finns granskningsmallar, en för varje studietyp (randomiserade kontrollerade studier, observationsstudier, diagnostiska studier, kvalitativa studier och systematiska översikter). Beroende på hur välgjord en studie är kan den få omdömet hög, medelhög eller låg studiekvalitet. Studier med hög och medelhög studiekvalitet utgör underlag för att syntetisera data och bedöma evidensstyrka.

Vid denna, mer noggranna granskning, visar det sig ofta att ytterligare några studier inte uppfyller kriterierna och därmed ska exkluderas (och föras in i exklusionslistan).

En viktig del av rapporten är tabeller med data från de studier som utgör det vetenskapliga underlaget. Tabellerna ska innehålla uppgifter om författare, population, intervention, kontroll, resultat och studiekvalitet. Det finns tabellmallar för frågor om intervention respektive för diagnostik. Tabellerna skrivs på engelska. Anledningen är att de, tillsammans med den engelska sammanfattningen, sprids internationellt via olika databaser. Om det saknas studier med medelhög eller hög studiekvalitet tabelleras data från studier med låg studiekvalitet. Tabellen används för att sammanfatta kunskapsläget men studierna kommer inte att vara tillräckligt underlag för syntes eller evidensstyrka.



Figur 2.2 Urvalsprocessen.

Syntes eller sammanvägning av resultat (Kapitel 9)

Nästa steg i processen är att syntetisera resultaten från studierna som ingår i det vetenskapliga underlaget, genom att t ex beräkna en effektstorlek. Om det finns flera studier är det lämpligt att undersöka om data går att väga samman i en metaanalys. Om metaanalysen visar att studierna är heterogena redovisas syntesarbetet rent deskriptivt. Metaanalysens så kallade ”forest plot” är användbar för såväl interventionsstudier som studier om diagnostisk säkerhet.

Metaanalyser kan utföras med hjälp av Cochrane Collaborations program RevMan som är tillgängligt kostnadsfritt.

Evidensgradering av resultaten (Kapitel 10)

Tillförlitligheten i de sammanvägda resultaten uttrycks med hjälp av en evidensstyrka. SBU använder evidensgraderingssystemet GRADE. GRADE är utarbetat av en internationell expertgrupp och systemet används i ökande utsträckning av organisationer och myndigheter, som t ex WHO, NICE och Cochrane Collaboration.

GRADE bygger i princip på erfarenheter från tidigare system men betonar i högre utsträckning patientnytta och risker. Evidensgraderingen beror av studiekvalitet, samt av hur tillförlitligheten påverkas av faktorer som heterogenitet i materialet, problem med relevans, statistisk osäkerhet och publikationsbias. GRADE har fyra nivåer: hög, måttlig, låg och mycket låg evidens. Resultat med låg respektive mycket låg evidens enligt SBU:s definition motsvaras av begränsat och otillräckligt vetenskapligt underlag.

Hälsoekonomi (Kapitel 11)

För att göra en allsidig utvärdering måste man bedöma metodens kostnadseffektivitet och de ekonomiska konsekvenserna av att metoden införs, utökas, minskas eller avvecklas.

Etiska och sociala aspekter (Kapitel 12)

Förutom att bedöma metodens effekter, risker och dess kostnadseffektivitet så bör utvärderingen inkludera etiska och sociala konsekvenser av metoden.

3. Strukturera och avgränsa översiktens frågor

VERSION 2011:1.1

Den första delen av projektet består i att besluta dels vilka frågor som ska besvaras inom projektets ram, dels projektets avgränsningar.

Projektgruppens uppdrag är ofta generellt hållet och behöver preciseras i ett begränsat antal frågor. Det gäller att välja ut de viktigaste. Ibland behöver projektgruppen konsultera andra intressenter för att säkerställa att de väsentligaste frågorna ringats in. Exempel på sådana intressenter kan vara de som föreslagit projektet liksom beslutsfattare, experter och patienter som inte deltar i utvärderingen.

Den strukturerade frågan

Till att börja med måste frågan struktureras, och ett vanligt sätt att göra det är att dela upp den i dess meningsbärande element. För de flesta medicinska frågor är PICO-systemet användbart. PICO är en förkortning för ”patient/population/problem, intervention/index test, comparison/control” (jämförelseintervention) och ”outcome” (utfallsmått) [1]. Genom att definiera vilka studiepopulationer, behandlingar, kontrollbehandlingar och utfallsmått som är relevanta för frågan fastställer man de huvudsakliga inklusionskriterierna. Dessa kompletteras av andra aspekter, som behandlingstid/uppföljningstid och lämplig studiedesign.

En väl strukturerad PICO ger ofta större chans till en mer specifik litteratursökning, vilket leder till färre artikelabstrakt som behöver sorteras bort. Faktaruta 3.1 sammanfattar PICO för interventionsstudier.

Det kan ofta löna sig att vara detaljerad när man ställer upp frågans inklusionskriterier. Då minskar risken för överinklusion – att man pga osäkerhet beställer artiklar som ligger utanför frågan – vilket leder till merarbete i ett senare skede.

Population

Vilka populationer är vi intresserade av? Populationen kan behöva specificeras mer eller mindre noggrant. Ibland räcker det med en diagnos, men i andra fall kan det behövas hög detaljnivå, t ex diagnos med svårighetsgrad av sjukdom, ålder, kön, etnicitet. I en välgjord studie är studiepopulationen definierad dels med tydliga inklusions- och exklusionskriterier, dels med tydligt redovisade baslinjedata. Ibland är populationen heterogen, och omfattar andra subgrupper än den aktuella populationen. Man kan behöva ta ställning till hur stor andel av studiedeltagarna som ska utgöras av den intressanta populationen, om inte resultaten är särredovisade för gruppen ifråga.

Faktaruta 3.1 Frågeställningens olika komponenter.

Population/ deltagare	Intervention/ metod	Jämförelsemetod/ kontroll	Effektmått
Här definieras den population som ska ha studerats. T ex: • ålder • kön • diagnos • sjukdomsgrad • riskfaktorer • övriga sjukdomar	Definition och beskrivning av metoden	Definition och beskrivning av jämförelsemetoden • Annan behandling • Placebo	Effektmått av direkt betydelse för individen såsom överlevnad, livskvalitet, sjuklighet och förändring av symtom Effektmått kan även vara komplikationer och biverkningar till följd av interventionen I hälsoekonomiska studier är effektmåttet ofta uttryckt i kostnad per kvalitetsjusterat vunnet levnadsår (QALY)

Som exklusionskriterium kan man här också specificera populationer som uppfyller inklusionskriterierna men som av olika skäl inte ingår i frågan. Det kan t ex röra sig om subgrupper av den intressanta populationen som utmärks av särskilda omständigheter, t ex samsjuklighet och medicinering.

Intervention

Vilka interventioner är vi intresserade av? Även här varierar detaljnivån beroende på fråga och den valda populationen. Man måste ibland specificera t ex dos, beredningsform och administrationssätt.

För frågor om t ex riskfaktorer för sjukdom kan exponering vara en lämpligare term än intervention.

Om frågan gäller diagnostisk säkerhet ska istället den experimentella metoden (indexmetoden) definieras här (Kapitel 7).

Jämförelseintervention

Vilken eller vilka åtgärder i jämförelsegruppen är acceptabla? Val av jämförelseintervention kan ofta vara avgörande för en studies relevans. Det är t ex vanligt att läkemedel som inte finns registrerade i Sverige används som jämförelseintervention. Ibland kan dock sådana substanser ändå vara relevanta, t ex om de kan anses vara representativa för en läkemedelsgrupp, exempelvis betablockerare eller bensodiazepiner. Motsatsen före-

kommer också, att man har valt ett registrerat läkemedel som kontroll, men ett som inte är representativt för gruppen. Ett exempel är studier av blodtrycksmediciner som använder betablockeraren atenolol som jämförelseintervention [2].

Ibland har doser av intervention och jämförelseintervention valts för att framhäva den terapeutiska effekten av interventionen eller för att tona ned risken för biverkningar [3]. För läkemedelsstudier förekommer det också att studien är designad på ett sätt som av farmakokinetiska skäl missgynnar kontrollsubstansen. Det kan t ex vara så att båda läkemedlen administreras peroralt, fastän upptaget av kontrollsubstansen är lågt [4].

Även utanför läkemedelsområdet är dessa faktorer av betydelse. För psykologiska studier kan dosen mätas i t ex antal, frekvens och längd på sessionerna. Korrekt administrerad kognitiv beteendeterapi utförs av personal med adekvat utbildning, men det förekommer i studier att KBT utförs också av outbildade vårdgivare. Det är ovanligt med placebointerventioner i psykologisk forskning. För psykologiska och sociala interventioner är det vanligt att jämförelsegruppen erbjuds ”sedvanlig vård”. Sådan vård kan variera väldigt mycket, och kan vara irrelevant för svenska förhållanden. Ibland kan också den jämförande psykologiska metoden vara sämre än ingen behandling alls, alltså direkt skadlig, vilket kan introducera en orättvis fördel för den undersökta interventionen [5].

För frågor om diagnostisk säkerhet definieras här den referensstandard som indextestet ska jämföras mot (Kapitel 7).

Effektmått

Vilka utfallsmått är lämpliga för att bedöma effekten av en åtgärd? I första hand bör man välja mått som är relevanta för patienten, såsom dödlighet, sjuklighet, lidande, funktionsnedsättning och livskvalitet. I andra hand kan man välja surrogatmått, alltså mätbara faktorer som i någon mån är relaterade till utfall som är relevanta för patienten. Exempel på surrogatmått är blodfetter, blodtryck och bentäthet.

Kompositmått är vanligt förekommande i klinisk forskning. Principen är att man genom att räkna samman flera olika effektmått kan få högre statistisk styrka i studien. Man bör dock vara försiktig med kompositmått, särskilt när surrogatmått ingår som en parameter. Ofta kan en statistiskt säkerställd effekt på ett kompositmått förklaras helt av effekt på ett surrogatmått eller en mindre viktig variabel som är relevant för patienten. Ibland kan kompositmåtten t o m maskera en negativ effekt av behandlingen på viktiga utfall som död och hjärt- och kärlhändelser [6].

För frågor om diagnostisk säkerhet är utfallsmåtten oftast sensitivitet och specificitet, mått som inte har direkt värde för patienten (Kapitel 7).

Behandlings- och uppföljningstid

Behandlings- och uppföljningstid måste ofta också anges i den strukturerade frågan. För behandling av kroniska tillstånd kan det t ex vara irrelevant att beakta korttidsstudier. Detsamma gäller frågor om prevention.

Studiedesign

Olika studiedesigner kan vara mer eller mindre lämpliga för att besvara en fråga (Faktaruta 3.2). Frågor om behandling besvaras t ex bäst med en randomiserad kontrollerad studie. Även frågor om diagnostisk säkerhet besvaras bäst med en randomiserad studie. Om projektgruppen bedömer att det redan finns många randomiserade studier kan det alltså vara ett skäl att avgränsa studietypen till att enbart granska randomiserade studier. För t ex nyare metoder och de flesta diagnostiska studier kommer man dock troligen behöva acceptera även studietyper som inte ger lika tillförlitliga resultat. Frågor om sällsynta biverkningar, eller om riskfaktorer för sjukdom, besvaras bäst med kontrollerade, prospektiva observationsstudier (t ex kohortstudier).

Faktaruta 3.2 Vanliga studiedesigner för olika frågor. Studiedesign med högst tillförlitlighet för varje fråga står först.

Frågan avser	Studietyp
Terapi/behandling/profylax	Randomiserad kontrollerad studie (RCT), kontrollerad studie
Prognos	Kohort
Biverkan/orsakssamband	RCT, kohort, fall–kontroll
Diagnos	Diagnostisk träffsäkerhetsstudie, RCT
Screening	RCT, tvärsnitt, kohort
Ekonomi	Kostnadseffektivitetsanalys
Etiologi	Kohort, fall–kontroll

Andra inklusionskriterier

Utöver PICO, behandlings- och uppföljningstid samt studiedesign kan man behöva definiera ytterligare inklusionskriterier.

Det kan ofta vara värdefullt att definiera i vilken miljö (”setting”) som studierna ska vara genomförda. Exempel på miljöer är akutmottagningar, arbetsplatser eller skolmiljö.

Ibland kan det vara nödvändigt att begränsa den studerade litteraturen till studier med en angiven minsta storlek på studiepopulationen. Sådana begränsningar bör om möjligt föregås av en analys av statistisk styrka.

Höga bortfall kan försvåra tolkningen av en studies resultat, eftersom det ofta är oklart vad som är skäl till varför personer väljer att avbryta deltagandet. Anledningar kan t ex vara utebliven effekt eller biverkningar. Höga bortfall är vanligt särskilt vid livsstilsstudier, där interventionen kräver mer än att bara ta en tablett. Bortfallet ökar också med tiden och det kan vara rimligt att ställa olika krav beroende på vilken uppföljningstid som har valts.

Andra avgränsningar

I praktiken behövs oftast ytterligare avgränsningar. De vanligaste är språk och publikationsdatum.

Språk

Utan språkbegränsningar kommer sökningen att omfatta studier på andra språk än engelska. Avgränsningar i språk görs dels med hänsyn till språkkunskaper i expertgruppen, dels om det är angeläget att beakta litteraturen på ett visst språk. Många alternativmedicinska studier är t ex publicerade på kinesiska, tyska och italienska, medan många kirurgistudier är publicerade på tyska.

Publikationsdatum

I vissa projekt kan det vara rimligt att avgränsa sökningen i tiden. Vissa metoder har t ex modifierats så mycket med tiden att det inte är relevant att läsa effektstudier på äldre versioner av metoden. Avgränsning på publikationsdatum kan också vara användbart vid uppdatering av tidigare rapporter.

Referenser

1. Boudin F, Nie JY, Bartlett JC, Grad R, Pluye P, Dawes M. Combining classifiers for robust PICO element detection. BMC Med Inform Decis Mak 2010;10:29.

2. Carlberg B, Samuelsson O, Lindholm LH. Atenolol in hypertension: is it a wise choice? Lancet 2004;364:1684-9.

3. Safer DJ. Design and reporting modifications in industry-sponsored comparative psychopharmacology trials. J Nerv Ment Dis 2002;190:583-92.

4. Johansen HK, Gotzsche PC. Problems in the design and reporting of trials of antifungal agents encountered during meta-analysis. JAMA 1999;282:1752-9.

5. Moos RH. Iatrogenic effects of psychosocial interventions for substance use disorders: prevalence, predictors, prevention. Addiction 2005;100:595-604.

6. Ferreira-González I, Permanyer-Miralda G, Busse JW, Bryant DM, Montori VM, Alonso-Coello P, et al. Methodologic discussions for using and interpreting composite endpoints are limited, but still identify major concerns. J Clin Epidemiol 2007;60:651-62.

4. Litteratursökning

VERSION 2012:1

Introduktion

I Kapitel 1 av SBU:s metodbok beskrivs de särskilda principer som kännetecknar arbetet med den systematiska översikten och som syftar till att minimera riskerna för att slump och godtycklighet påverkar översiktens slutsatser. En av dessa principer är den systematiska litteratursökningen. Målet är att om möjligt fånga alla för frågeställningen relevanta studier. Det här kapitlet handlar om litteratursökningen som en del av projektprocessen med fokus på sökning av originalartiklar i internationella ämnesdatabaser med vetenskapligt innehåll.

I arbetet med att försöka fånga alla relevanta studier används också kompletterande metoder. Sökningar görs oftast i flera olika ämnesdatabaser och vid behov görs även sökningar i citeringsdatabaser. Citeringssökning innebär att man utgår från en bestämd forskare eller en artikel för att ta reda på om den är citerad och i så fall av vem. Exempel på citeringsdatabaser är Web of Science och Scopus. Andra kompletterande metoder är att analysera redan funna studiers referenslistor, ofta kallad kedjesökning, samt det självklara att experters kunskap om ämnesområdet noggrant tas till vara. Många HTA-organisationer, som t ex Cochrane Collaboration [1], anger att handsökning av specifika tidskriftstitlar utförs och att sökning av så kallad grå litteratur ingår i arbetet med systematiska översikter [2]. Handsökning innebär att vissa för frågeställningen viktiga tidskrifter söks igenom sida för sida vilket är mycket tidskrävande. De två sistnämnda metoderna används sällan på SBU. Vissa studier pekar på att sökning av grå litteratur i form av konferensabstrakt visserligen minimerar publikationsbias (endast hälften av innehållet av dessa abstrakt resulterar i vetenskapliga artiklar) men samtidigt är informationen ofta alltför knapp vad gäller metoddelen. [3]

Litteratursökningen – en del av projektprocessen

Arbetet med att skapa en så heltäckande sökstrategi som möjligt är ett samarbete mellan informationsspecialist, projektledare och projektets experter.

Några framträdande drag i processen kan urskiljas: förberedande sökningar, testsökning, huvudsökning och mot projektidens slut en uppdaterande sökning. Utgångspunkten för litteratursökningen är alltid uppdragets frågeställning som struktureras i en projektplan. Fördelen med att samarbeta redan från start i arbetet med projektplanen är att informationsspecialistens arbete med sökstrategin effektiveras genom den ökade förståelse för frågans olika aspekter som denna får. En annan lika viktig aspekt är att informations-

specialistens kunskap och erfarenheter av att omsätta en frågeställning till en sökstrategi kan bidra till att strukturera frågan. Information om för frågeställningen lämpliga databaser kan tidigt presenteras.

Före projektstart

Innan ett projekt startar bör ett förberedelsearbete ha gjorts för att kontrollera om liknande projekt är under arbete i någon annan HTA-organisation eller om andra aktuella systematiska översikter redan finns. Viktiga databaser här är Cochrane Librarys deldatabaser Cochrane Reviews och DARE.

Faktaruta 4.1 Databaser/webbsidor som innehåller systematiska översikter och HTA-rapporter.

- **Cochrane Database of Systematic Reviews (CDSR)**
  Innehåller flera deldatabaser, bl a Cochrane Database of Systematic Reviews
  www.thecochranelibrary.com

- **HTA-nätverkets databas**
  Svenska HTA-rapporter
  www.sbu.se/HTADatabas

- **Nasjonalt kunnskapssenter for helsetjenesten**
  Nationell HTA-organisation (Norge)
  www.kunnskapssenteret.no

- **National Institute for Health Research Evaluation, Trials and Studies Coordinating Centre**
  Nationell HTA-organisation (Storbritannien)
  www.hta.ac.uk/about

- **Canadian Agency for Drugs and Technologies in Health**
  Nationell HTA-organisation (Kanada)
  www.cadth.ca

Testsökning

När det är beslutat att ett projekt ska starta formulerar informationsspecialisten, i samarbete med projektledaren, sökstrategier för testsökningar som sedan utförs. Testsökningarna syftar till att undersöka bl a följande frågor:

- Hur är relevanta studier indexerade och vilka termer förekommer i titel och abstrakt?
- Är vår frågeställning tillräckligt väldefinierad, eller behöver den förtydligas ytterligare?
- Hur stora sökmängder kan vi förvänta oss?

Projektets experter har en mycket viktig roll genom att förse informationsspecialisten med ”kärnartiklar” som används för att utveckla sökstrategierna och överhuvudtaget bidra med kompletterande litteratur som inte fångas av informationssökningarna. Experterna kan också bidra med begrepp och uttryck hämtade från det aktuella ämnesområdet och bedöma om sökresultatet är passande för projektets fråga(or) eller om korrigeringar av sökstrategin bör göras.

Formerna för det konkreta samarbetet mellan informationsspecialist och experter kan genomföras på olika och ibland kompletterande sätt. Det kan vara i form av möten, fysiska såväl som onlinemöten, där sökstrategierna och sökresultatet diskuteras. Experterna kan också ges möjlighet att själva bläddra i det preliminära sökresultatet, genom t ex Collections från PubMed eller i form av ett bibliotek från ett referenshanteringssystem, och kan därefter meddela sina synpunkter till informationsspecialisten. Projektledarens roll i detta samarbete kan variera, men det är viktigt att denne är väl insatt i hur arbetet fortskrider.

”Huvudsökning”

När sökstrategin är genomarbetad utförs sökningen i den databas som valts som första databas. För projekt och frågeställningar inom medicinområdet görs testsökningar och sedan även huvudsökningarna först i PubMed, men för frågeställningar inom andra områden kan andra databaser vara förstahandsvalet.

Nästa steg är att anpassa sökstrategin till resterande databaser. Sökstrategier och sökningar inom hälsoekonomi och etikområdet formuleras och utförs. Förutom att söka i de databaser som finns i projektplanen, kan man till hälsoekonomiområdet även söka i databasen NHS EED (NHS Economic Evaluation) som ingår i Cochrane Library samt i någon ekonomisk ämnesdatabas som t ex EconLit. Alla sökstrategier och sökresultat dokumenteras noggrant som en viktig del av SBU:s krav på hög tillförlitlighet och transparens.

Sökresultaten importeras till ett referenshanteringsprogram där dubblettkontroll görs. När alla sökningar är gjorda och alla dubbletter borttagna återstår den manuella granskningen av de framsökta abstrakten. Sökningen identifierar ett antal referenser, men arbetet med att avgöra hur relevanta de är i förhållande till frågeställningen måste sedan göras manuellt.

Uppdateringssökning

Om det har gått lång tid sedan huvudsökningarna gjordes till dess att rapporten publiceras, bör en uppdateringssökning göras innan rapporten publiceras. Denna sökning görs för att identifiera de allra senaste publicerade studierna och hinna få med dem i arbetet.

4

Utformning av sökstrategi

Som beskrivits i tidigare kapitel är en väl strukturerad och definierad frågeställning av avgörande betydelse för en effektiv litteratursökning. Att strukturera frågeställningen innebär helt enkelt att den delas upp i sina olika beståndsdelar och att varje del analyseras. Och att de beslut som tas dokumenteras i projektplanen.

Från PICO till sökning

Som en hjälp i arbetet med att strukturera frågeställningen används för studier om interventioner och diagnostik en så kallad PICO (”population, intervention, comparison/control, outcome”) och för studier som bygger på kvalitativ data kan projektets frågor struktureras med hjälp av en så kallad SPICE (”setting, perspective, intervention, comparison, evaluation”).

Arbetet med att strukturera frågeställningen innebär att projektets inklusions- och exklusionskriterier tar form. I dessa ingår också att ta beslut om eventuella avgränsningar, t ex begränsning till en viss tidsperiod, avgränsningar till vissa språk eller vissa studieupplägg. Allt detta har betydelse för hur sökstrategin utformas. Sökstrategin utgår från frågeställningens PICO, men det är viktigt att uppmärksamma att det inte betyder att alla delar av en PICO/SPICE alltid ska vara med som en del av sökstrategin.

”Building block strategy”

När man formulerar en sökstrategi använder man sig vanligtvis av det som på engelska brukar kallas ”building block strategy” och som på svenska kan kallas blocksökning. Varje del av PICO som man valt att använda i sökstrategin, motsvaras vanligtvis av ett block av söktermer och sökfraser. Ibland kan vissa delar av frågeställningen motsvaras av två block i sökningen. Om frågeställningen t ex handlar om populationen ”äldre personer med urininkontinens” motsvaras detta förslagsvis av två block; ett block för äldre personer och ett block för urininkontinens. Varje block söks för sig för att sedan kombineras med varandra för ett slutgiltigt sökresultat.

Booleska operatorer för att kombinera sökord

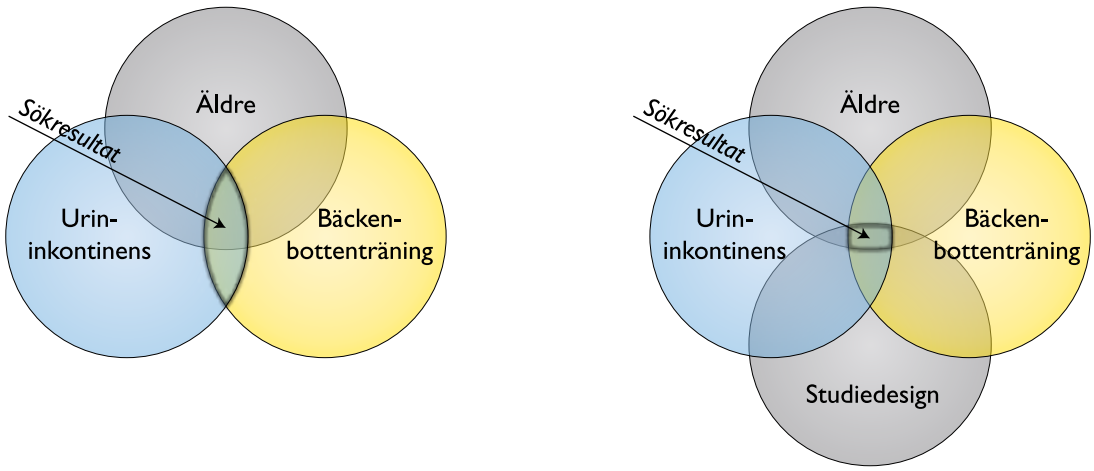
De enskilda blocken av söktermer som ska ingå i sökningen skapas genom att kombinera begrepp och termer med en boolesk operator. De booleska operatorerna AND, OR, NOT är programmerade att ge databasen specifika instruktioner och ska inte blandas

samman med ordens vardagliga betydelse. Inom varje block av söktermer kombineras synonyma begrepp och andra näraliggande termer med den booleska operatorn OR. Operatorn OR mellan varje sökterm inom ett block ger databasen instruktionen att söka antingen det ena eller andra söktermen eller alla i blocket förekommande termer. Genom att använda OR garderar man sig för den mångfald av olika uttryck som kan användas i t ex ett abstrakt för en och samma sjukdom, intervention etc. Sökresultatet utvidgas i jämförelse om man bara sökt med ett sökord.

När varje block av sökord är sökta kombineras de med ett booleskt AND mellan blocken. Instruktionen till databasen är då att minst *ett* ord ur varje block måste finnas i varje referens av sökresultatet. Nu specificeras sökningen och sökresultatet snävas in.

Ett smidigt sätt att kombinera block är att använda respektive databas funktion för sökhistorik.

Den booleska operatorn NOT används för att ge databasen instruktionen att något inte ska förekomma i sökresultatet. Vanligen används denna med stor försiktighet.



Figur 4.1 Sökresultat med den booleska operatorn AND.

Parentessökning

Parenteser används i en sökstrategi där olika booleska operatorer ingår för att bestämma i vilken ordning databasen ska söka söktermerna och operatorerna.

Exempel: incontinence AND (urine OR urinary OR stress OR urge)

Parentesen ger databasen instruktionen att börja med att utföra sökningen inom parentesen. Det sökresultatet kombineras sedan med söktermen ”incontinence” och ett booleskt AND.

Olika typer av sökord – indexeringsord

En sökstrategi till en systematisk översikt består av en blandning av indexeringsord och fritextord, för att fånga så många av de relevanta studierna som möjligt.

Indexeringsorden hämtas från den särskilda alfabetiskt hierarkiskt uppställda ordlista, tesaurus, som varje stor internationell ämnesdatabas har. Medlines (PubMed) tesaurus kallas MeSH medan PsycInfos kallas Thesaurus of Psychological Index Terms. Eftersom olika databasers tesaurus använder olika begrepp och uttryck, olika indexeringsord eller kontrollerade sökord, måste alla sökstrategier omformuleras och anpassas till varje specifik databas.

Huvuddelen av alla artiklar som läggs in i en databas indexeras, dvs en indexerare lägger till ett antal termer ur tesaurus till varje artikel. Dessa indexeringsord ska beskriva innehållet i en artikel och ibland även studiedesign, publikationstyp m m. En tesaurus syftar till att försöka skapa ett enhetligt sätt att benämna innehållet i en databas samtidigt som den skapar relationer mellan begreppen i det hierarkiska systemet.

Fritextord

Den andra typen av sökord som används kallas fritextord. Det är söktermer valda för att matcha ord som förekommer i databasens beskrivning av varje specifik studie. Här kan man bestämma var i beskrivningen orden får förekomma. Det är vanligt att en begränsning görs så att fritextsorden matchar ord som finns i referensernas titlar och abstrakt.

Fördelar och nackdelar med indexeringsord respektive fritextord

Fördelar med att söka med hjälp av databasernas indexeringsord är att de är enhetliga, varje referens får ett antal distinkta termer som syftar till att beskriva artikelns innehåll. Ett abstrakt ska också beskriva en artikels innehåll men att söka på ord i beskrivande text kan leda till irrelevanta träffar. Med indexeringsord behöver man inte heller ta hänsyn till synonymer och stavningsvarianter som man måste göra med fritextord. En nackdel kan vara att de ibland blir för generella för att passa den aktuella frågeställningen. Viktigt är också att val av titel och hur abstrakt skrivs har betydelse för hur artikeln kommer att indexeras och naturligtvis måste den mänskliga faktorn vad gäller felindexering beaktas.

Fördelar med fritextord är att man med hjälp av dessa även hittar studier som ännu inte blivit indexerade. Det betyder att för att fånga de allra senaste publicerade artiklarna i t ex den viktiga PubMed så räcker det inte att söka med indexeringsord. En kombination med fritexttermer behövs. Fritexttermer kan också vara till hjälp när databasens indexeringsord är för generellt för att passa den aktuella frågeställningen.

Avgränsningar

När frågeställningens PICO arbetas fram tar man också ställning till vilka avgränsningar som frågan ska ha och om dessa ska ingå i sökstrategin eller sållas manuellt vid granskningen av abstrakt.

Avgränsningar kan gälla populationens ålder, kön, språk, begränsningar i tid eller studiedesign etc.

Internationella databaser har inbyggda funktioner, Limits, för avgränsningar. I en del databaser, som t ex PubMed, är användandet av vissa Limits liktydigt med att söka med MeSH-termer vilket betyder att man inte får träff på nya artiklar som ännu inte är indexerade med MeSH. Det gäller bl a funktionerna Ages, Article Type och Species. Andra avgränsningar som språk och tid är inte kopplade till MeSH utan man får träff även på oindexerade artiklar.

Alla beslut om avgränsningar tas i projektgruppen

Experternas kunskaper om forskningsområdets utveckling har stor betydelse för vilka avgränsningar som är lämpliga tillsammans med SBU:s personals kunskaper och erfarenheter av metoderna för att skapa översikter. Hänsyn måste också tas till projektens tidsramar och resurser.

Språk: I databaserna kan avgränsningar till olika språk lätt göras. Projektgruppen måste besluta om det är av intresse att få en överblick över även icke-engelskspråkiga studier (abstrakten är alltid på engelska men inte själva studien) eller om begränsningar ska göras.

Tidsperiod: Det kan finnas skäl att ange en begränsad tidsperiod i sökstrategin. Vid uppdateringar av tidigare sökningar kompletteras den tidigare sökningen.

Studiedesign: Projektgruppen måste också ta beslut om studiedesign ska ingå i själva sökstrategin eller bara som inklusionskriterier som hanteras i abstraktgranskningen.

Sökfilter (på engelska ”search filters” eller ”hedges”) är ett hjälpmedel för att underlätta sökningen en viss typ av studier t ex studiedesign. Moderna sökfilter är validerade, dvs kontrollerade för att hitta så många relevanta studier som möjligt samtidigt som antalet icke relevanta studier som fångas ska begränsas. Sökfiltren är anpassade till både olika versioner av en databas, t ex Medline, och till helt olika databaser. Sökfiltret kombineras med sökstrategins övriga block.

Vid Centre for Reviews and Dissemination (CRD) samlas, utvärderas och publiceras sökfilter [4].

Några sammanfattande punkter att beakta vid utformning av en sökstrategi till en systematisk översikt:

- Skapa sökblock som består av både indexeringsord och fritextord.
- Sök på så få delar av PICO som möjligt, sålla resten vid abstraktgranskningen.
- I vissa frågeställningar motsvarar *en* del av PICO flera sökblock.
- Det är oftast populationen samt interventionen som är lämpliga att söka på.

Litteratursökningens omfattning: en balansgång

Förhoppningen är att systematiska litteraturöversikter baseras på all existerande relevant litteratur. Den optimala litteratursökningen till ett sådant projekt vore därför en sökning som både hittar alla relevanta studier och ingenting annat än de relevanta studierna, dvs en sökning med 100 procents precision. När man gör en litteratursökning kan man ha olika ansatser, man kan göra sökningen mer eller mindre omfattande; ”bred sökning” och ”smal sökning”. Hur omfattande sökningen görs är ofta i praktiken beroende av hur många träffar som sökningen genererar, eftersom det alltid är en människa som går igenom resultatet av sökningen (i form av listade abstrakt).

Faktaruta 4.2 Begrepp som används vid beskrivning av sökresultatet [5].

Precision = Andelen relevanta hittade artiklar i proportion till det totala antalet hittade artiklar.

”Recall” = Andelen av de relevanta träffarna som man hittade i förhållande till det totala antalet relevanta artiklar.

Smal sökning

En litteratursökning där man exempelvis söker efter två ord i artikelns titelfält och kombinerar dessa med ett booleskt AND, ger naturligtvis få träffar och de träffar man får är antagligen till stor del relevanta. Samtidigt har man säkerligen missat stora delar av den relevanta litteraturen eftersom man inte tagit hänsyn till varierande terminologi. En smal sökning har alltså ofta hög precision, men det kan naturligtvis hända att den smala sökningen inte alls träffar ”mitt i prick” utan snarare helt utanför. En smal sökning brukar inte vara tillräckligt omfattande för en översikt men fyller en funktion vid litteratursökning för andra ändamål.

Bred sökning

I arbetet med en systematisk översikt bör man sträva mot att göra en bred sökning, en sökstrategi som tar hänsyn till varierande indexering, bristande indexering och att vissa

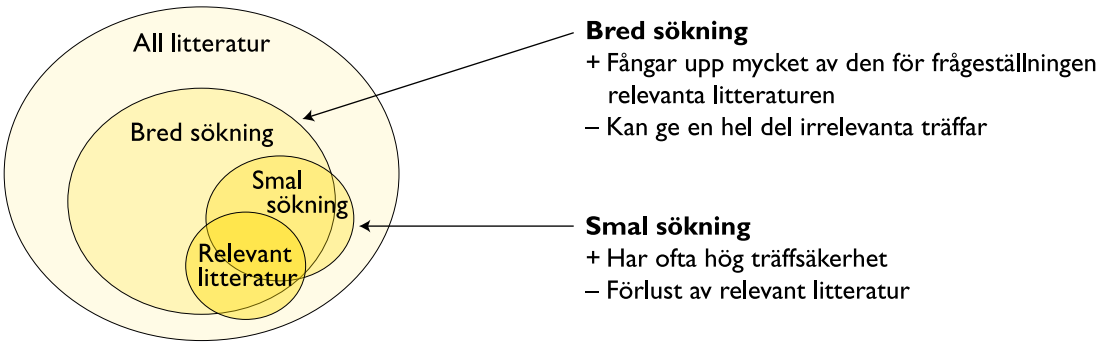
studier inte blivit indexerade. Syftet med en bred sökning är att ha hög ”recall”, att hitta så många som möjligt av de existerande studierna som svarar på frågeställningen. Naturligtvis vet man inte när sökningen görs hur stor andel av de relevanta studierna som faktiskt kommer att fångas, men en bred sökning ökar möjligheterna att finna det mesta. Nackdelen är att ju bredare sökning desto fler irrelevanta träffar kommer sökningen att fånga.

Det finns många olika sätt att bredda eller smalna av litteratursökningen, och några tips återges i Faktaruta 4.3.

Faktaruta 4.3 Tips för en smalare respektive bredare sökning.

Tips för en smalare sökning	Tips för en bredare sökning
• Använd endast indexeringstermer. • Avgränsa indexeringsorden med funktioner för huvudämne och aspektord (i PubMed: ”Major Topic” respektive ”Subheadings”). • Begränsa sökningen till publiceringstid, språk, åldersgrupp. • Om varje del av en PICO motsvaras av ett block med söktermer blir sökresultatet smalare ju fler delar som ingår i sökstrategin med ett booleskt AND mellan varje block. • När du söker med fritextord, sök endast efter ord i referensernas titlar. • När du söker med fritextord, sök på specifika ord eller fraser (t ex ”cost-effectiveness” i stället för ”cost*” och ”qualitative study” i stället för ”qualitative”). • Undvik att söka på förkortningar om samma förkortning kan betyda olika saker.	• Sök med både indexeringsord och fritextord. • Ta hänsyn till att en tesaurus är ett föränderligt hjälpmedel. Det kan finnas olika sätt att indexera samma sak eller näraliggande företeelser. • Sök i flera för ämnesområdet relevanta databaser. • Sök med få block (ofta block för population AND intervention). • Lägg till alternativa stavningar och böjningsformer för fritextorden. • Trunkera fritextorden när det är tillämpligt, dvs sök på ordstam som slutar med ett trunkeringstecken (vanligen *). Men kontrollera och avstå om trunkeringen ger för många irrelevanta träffar.

Skillnaden mellan breda och smala sökningar åskådliggörs i Figur 4.2.



Figur 4.2 Skillnaden mellan breda och smala sökningar.

”Number needed to read”

Sökningens precision kan också uttryckas som ”number needed to read” (NNR), ett mått som tar antalet abstrakt att läsa i beaktande. NNR beskriver hur många abstrakt man måste läsa för att finna en relevant artikel (NNR=1/precisionen). NNR avgörs inte bara av hur bred/smal sökningen är utan också av forskningsområdets omfattning, samt av hur väl avgränsad frågeställningen är. Syftar projektet till att besvara en frågeställning där det finns få publicerade studier, så är det ganska okomplicerat att göra sökningen bred. En sådan sökning riskerar inte att missa särskilt många relevanta artiklar, samtidigt som arbetsbördan inte behöver bli särskilt stor för dem som sedan kommer att granska de funna referenserna.

Om projektet däremot vill besvara en frågeställning där det finns ett stort antal publicerade studier, ställs frågan om sökningens bredd på sin spets. Hur många abstrakt är projektets experter beredda att läsa igenom manuellt för att vara säkra på att ingenting missats?

Balansen mellan hur smal och hur bred man gör sökningen är alltså i mångt och mycket en fråga om tid, hur många personer som arbetar i projektet och var man lägger arbetsbördan. Ibland går det kanske snabbare och enklare att granska ett stort antal referenser i jämförelse med den tid det tar att snäva in sökningen på ett sätt som gör att man inte missar alltför många relevanta studier. Å andra sidan är alternativet med ett för stort antal söktträffar med högt NNR (dvs man måste läsa ett stort antal irrelevanta artiklar för att hitta en relevant) inte heller oproblematiskt. Den mänskliga faktorn gör att det kan vara svårt att hålla koncentrationen uppe vid granskning av ett stort antal abstrakt, och på så vis riskerar man också att relevanta studier sållas bort av misstag. Det kan dock påpekas att det inte behöver ta alltför mycket tid i anspråk att granska en abstraktlista, trots att antalet abstrakt vid första anblicken kan se ut att vara ett ohanterbart antal:

> *”At a conservatively-estimated reading rate of two abstracts per minute, the results of a database search can be ‘scan-read’ at the rate of 120 per hour (or approximately 1 000 over an 8-hour period)” [6].*

Val av databaser

Det finns flera studier som visar att det inte räcker att söka i endast en databas när syftet är att hitta alla studier som svarar på den aktuella frågeställningen [7]. Vilka databaser, och hur många databaser, som är lämpliga att söka i beror helt på frågeställningens ämne.

Enligt den mall som vanligtvis används vid kvalitetsgranskning av systematiska översikter, AMSTAR (Bilaga 6), krävs det utförlig sökning i minst två databaser för att sökningen ska bedömas vara tillräcklig. På SBU genomsöks vanligtvis minst tre databaser; för frågor inom medicinområdet räcker det ofta med sökningar i PubMed, Embase och i Cochrane Library. Om frågeställningen är av mer multidisciplinär karaktär, bör detta tas i åtanke vid val av databas.

Observera att även om en referens finns inkluderad i en databas innebär detta inte att den är lätt att hitta. Därför kan kompletterande sökningar i andra databaser vara av värde, eftersom samma referens kan vara indexerad på olika sätt i olika databaser.

Faktaruta 4.4 Exempel på bibliografiska databaser som är viktiga för systematiska litteraturöversikter inom hälso- och sjukvårdsområdet.

PubMed (www.ncbi.nlm.nih.gov/pubmed)
PubMed innehåller ca 22 miljoner referenser till artiklar och ett urval fulltextartiklar från mer än 5 000 biomedicinska tidskrifter (2012). Databasen ger en bred täckning inom hälso- och medicinområdet. PubMeds huvudsakliga innehåll utgörs av databasen Medline. Utmärkande för artiklarna i Medline är att de är indexerade enligt databasens särskilda tesuarus MeSH (Medical Subject Heading). Förutom dessa finns ett växande antal artiklar i PubMed som väntar på indexering och som känns igen genom kommentarerna ”PubMed – in process” alternativt ”Supplied by publisher”. Databasen produceras av National Library of Medicine i USA och är kostnadsfritt tillgänglig via internet.

Embase (www.embase.com)
Embase är den andra stora databasen inom medicinområdet. Embase innehåller ca 25 miljoner referenser från 7 600 tidskrifter (2012). I Embase finns möjlighet till en integrerad sökning med databasen Medline, men Embase innehåller inte PubMeds övriga innehåll eller MeSH-databasen. Embase har en utvecklad tesaurus, Emtree, som brukar framhållas som särskilt bra på farmakologi som är ett av databasens centrala ämnesområden. Förutom artiklar innehåller Embase även konferenshandlingar. Liksom i PubMed kan artiklar ”In process” sökas men här finns också ”Article in press”, dvs ännu inte publicerade artiklar. Embase produceras av det europeiska vetenskapliga förlaget Elsevier och innehåller ett större antal europeiska tidskrifter på respektive europeiskt språk än den amerikanska PubMed. Databasen är avgiftsbelagd.

Faktarutan fortsätter på nästa sida

Faktaruta 4.4 Fortsättning.

Cochrane Library (www.thecochranelibrary.com)
Cochrane Library består av flera olika deldatabaser. Förutom Cochrane Database of Systematic Reviews som innehåller de egna systematiska översikterna, finns bl a Cochrane Central Register of Controlled Trials (Central), Cochrane Methodology Register och NHS Economic Evaluation Database (NHS EED). Både NHS EED och DARE, som också är en deldatabas i Cochrane Library, produceras av CRD, Centre for Reviews and Dissemination. I databasen Health Technology Assessment Database samlar man nya och pågående projekt utanför Cochrane-samarbetet. SBU tillhandahåller Cochrane Library gratis för personer med svensk IP-adress.

CINAHL (www.ebscohost.com/biomedical-libraries/the-cinahl-database)
CINAHL (Cumulative Index to Nursing and Allied Health Literature) är en databas över artiklar om omvårdnad, sjukgymnastik, arbetsterapi etc. Den innehåller ca 2,3 miljoner referenser ur ca 3 000 tidskrifter. Databasen tillhandahålls av EBSCO och åtkomst är avgiftsbelagd.

PsycInfo (www.apa.org/psycinfo)
PsycInfo är en databas inom psykologi, beteendevetenskap och näraliggande ämnesområden. Databasen ger referenser till ca 3 miljoner vetenskapligt granskade artiklar ur ca 2 500 tidskrifter (2012), böcker och dissertationer. PsycInfo är avgiftsbelagd och produceras av American Psychological Association (APA).

Dokumentation

För att en databassökning ska gå att repetera är det viktigt att tillvägagångssättet dokumenteras. Sökdokumentationen bör sedan finnas tillgänglig för dem som läser den systematiska översikten. Det finns ingen allmän standard för hur sökdokumentationsmallen ska se ut, däremot bör följande information redovisas:

- databasens namn
- databasleverantörens namn
- datum när sökningen gjordes
- exakta söktermer och vilken typ av term det är, dvs indexeringsord eller fritext
- eventuella begränsningar
- hur termerna kombinerats.

SBU:s sökdokumentationsmall visas i Exempel 4.1.

Redogör också för eventuella komplementära sökmetoder om sådana använts för att hitta relevant litteratur till projektet, exempelvis handsökning eller kedjesökning.

Exempel 4.1 Sökdokumentation.

Pubmed via NLM 17 November 2011
Title: Pelvic floor muscle training as an intervention for elderly with urinary incontinence

	Search terms	Items found
Population: aged		
1.	"Aged"[Mesh:NoExp] OR "Aged, 80 and over"[Mesh] OR "Frail Elderly"[Mesh] OR Geriatrics[MeSH] OR Homes for the Aged[MeSH]	2 038 796
2.	(older patient*[TI] OR older adult[TI] OR older adults[TI] OR older women[TI] OR older men[TI] OR geriatric[TI] OR geriatrics[TI] OR elderly[TI] OR elders[TI] OR Vulnerable elder[TI] OR Vulnerable elders[TI] OR senior[TI] OR seniors[TI] OR community-dwelling[TIab] OR nursing home[TI] Or nursing homes[TI] OR care home[TI] OR care homes[TI] OR oldest old[TI] OR frail[TI]) NOT medline[SB])	7 972
3.	*1 OR 2*	*2 046 528*
Population: urinary incontinence		
4.	Urinary Incontinence[MeSH:NoExp] OR Urinary Incontinence, Stress[MeSH] OR Urinary Incontinence, Urge[MeSH] OR Nocturia[MeSH] OR Urinary Bladder, Overactive[MeSH] OR "Diurnal Enuresis"[Mesh] OR overactive bladder[tiab]	25 556
5.	(Mixed incontinence[tiab] OR Stress incontinence[tiab] OR Stress urinary[tiab] OR overactive bladder[tiab] OR bladder overactivity[tiab] OR bladder control[tiab] OR urge to void[tiab] OR (Incontinence[ti] AND (urine[ti] OR urinary[ti] OR stress[ti] OR urge[ti])) NOT medline[SB]	1 146
6.	*4 OR 5*	*26 393*
Intervention: pelvic floor muscle training		
7.	(Pelvis[MeSH:NoExp] OR Pelvic Floor[MeSH]) AND (Muscle Contraction[MeSH] OR Exercise Therapy[MeSH:NoExp] OR Physical Therapy Modalities[MeSH])	1 407
8.	pelvic muscles exercise*[tiab] OR Pelvic muscle exercise*[tiab] OR Bladder and pelvic muscle training[tiab] OR pelvic floor muscle training[tiab] OR pelvic floor re-education[tiab] OR pelvic exercise*[tiab] OR pelvic floor training[tiab] OR pelvic muscle precontraction[tiab] OR pelvic floor exercise*[tiab] OR pelvic muscle re-education[tiab] OR (pelvic floor[ti] AND (training[ti] OR exercise*[ti] OR education[ti]))	1 040
9.	*7 OR 8*	*1 972*
Combined sets		
10.	*3 AND 6 AND 9*	*350*

The search result, usually found at the end of the documentation, forms the list of abstracts.

*[MeSH] = Term from the Medline controlled vocabulary, including terms found below this term in the MeSH hierarchy; [MeSH:NoExp] = Does not include terms found below this term in the MeSH hierarchy; [MAJR] = MeSH Major Topic; [TIAB] = Title or abstract; [TI] = Title; [AU] = Author; [TW] = Text Word; Systematic[SB] = Filter for retrieving systematic reviews; * = Truncation; " " = Citation Marks, searches for an exact phrase*

Referenshantering

För att kunna hantera den stora mängd referenser som omfattas av arbetet med en systematisk litteraturöversikt krävs ett kraftfullt referenshanteringsprogram, t ex EndNote eller Zotero. Med hjälp av programmet importeras alla referenser från databassökningarna till ett bibliotek som är specifikt för frågeställningen eller hela projektet.

4

Referenser

1. Cochrane Collaboration. [2012; citerad 27 september 2012]. Tillgänglig från: http://www.cochrane.org/

2. Lefebvre C, Manheimer E, Glanville J. Chapter 6: Searching for studies. In: Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions. Version 5.1.0. [Uppdaterad mars 2011, citerad 27 september 2012]. Tillgänglig från: http://www.cochrane-handbook.org/

3. Dundar Y, Dodd S, Dickson R, Walley T, Haycox A, Williamson PR. Comparison of conference abstracts and presentations with full-text articles in the health technology assessments of rapidly evolving technologies. Health Technol Assess 2006;10:1-145.

4. Centre for Reviews and Dissemination (CRD): The InterTASC Information Specialists' Sub-Group Search Filter Resource [Citerad den 27 september 2012]. Tillgänglig från: http://www.york.ac.uk/inst/crd/intertasc/index.htm

5. Shariff SZ, Cuerden MS, Haynes RB, McKibbon KA, Wilczynski NL, Iansavichus AV, et al. Evaluating the impact of MEDLINE filters on evidence retrieval: study protocol. Implementation Sci 2010;5:58.

6. Lefebvre C, Manheimer E, Glanville J. Chapter 6.4.4: Sensitivity versus precision. In: Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions. Version 5.1.0. [Uppdaterad mars 2011, citerad 27 september 2012]. Tillgänglig från: http://www.cochrane-handbook.org/

7. Lefebvre C, Manheimer E, Glanville J. Chapter 6.1.1.2: Minimizing bias. In: Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions. Version 5.1.0. [Uppdaterad mars 2011, citerad 27 september 2012]. Tillgänglig från: http://www.cochrane-handbook.org/

5. Bedömning av en studies relevans

VERSION 2011:1.1

Genomgången av den insamlade litteraturen inleds med en bedömning av studiernas relevans för frågan, alltså hur väl de uppfyller de uppställda inklusionskriterierna. Syftet med relevansbedömningen är att sålla bort de studier som är irrelevanta för frågan. Endast studier som bedöms vara relevanta går vidare till kvalitetsgranskning.

Till skillnad från kvalitetsbedömning av studier graderas inte en studies relevans. En studie är alltså antingen relevant eller inte relevant för frågan. I samband med kvalitetsgranskningen av relevanta studier görs en bedömning av extern validitet eller generaliserbarhet, som ej bör förväxlas med relevans.

Relevansbedömningens två steg

Relevansbedömningen görs i två olika steg (Figur 2.2). Som tidigare angetts bör varje steg utföras av två oberoende granskare. Varje steg bör också dokumenteras noggrant.

I steg ett görs en grovsållning utifrån artiklarnas titlar och abstrakt (artikelsammanfattning). Vid tveksamhet är det i detta skede ofta bättre att fria än att fälla. Studier som av minst en av granskarna bedöms kunna vara relevanta beställs i fulltext.

I steg två granskas de beställda fulltextartiklarna med avseende på relevans (se granskningsmall, Bilaga 1). Studier som bedöms vara relevanta inkluderas i den systematiska litteratursammanställningen och går vidare till kvalitetsgranskning (Kapitel 6–7). Studier som i detta steg inte bedöms vara relevanta exkluderas, men såväl antal som orsak måste dokumenteras. Vid oenighet tas studien upp för diskussion i hela gruppen.

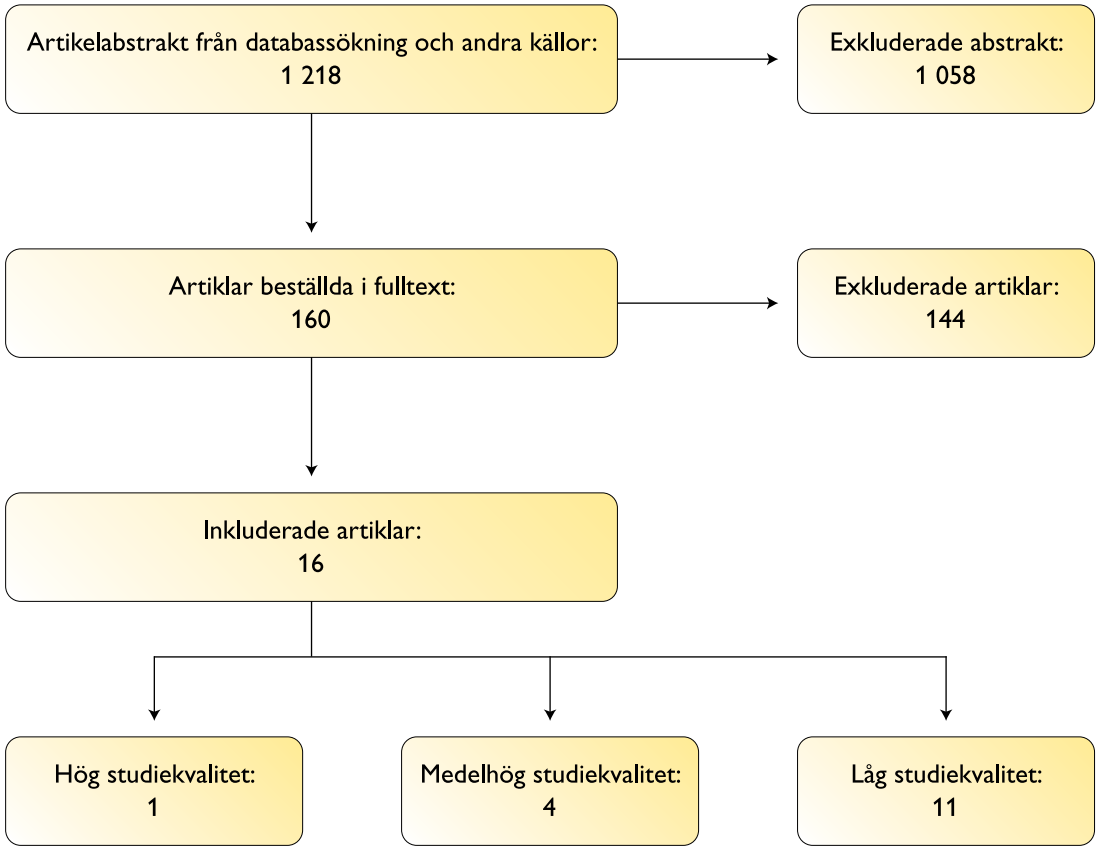
Principer för relevansbedömning

Bedömningen av en studies relevans utgår från de uppställda inklusions- och exklusionskriterierna (Kapitel 3). Vid bedömningen av en studies relevans stämmer man av de olika kriterierna mot de uppgifter som redovisas i studien.

Dokumentation av exkluderade studier

Urvalsprocessen redovisas i form av ett flödesschema, där antal artikelabstrakt och fulltextartiklar ska kunna spåras tillbaka till litteratursökningen (Figur 5.1). Med anledning av de höga krav som ställs på transparens vid framtagandet av en systematisk litteraturöversikt är det också viktigt att dokumentera skälet till att en artikel som lästs i fulltext

inte inkluderas. Ibland kan de vanligaste skälen också anges i flödesschemat. Fulltext-artiklar som exkluderats brukar redovisas i en bilaga till rapporten, med angivet skäl för exklusion. Några exempel redovisas i Faktaruta 5.1.



Figur 5.1 Exempel på flödesschema som redovisar antal inkluderade/exkluderade artiklar.

Faktaruta 5.1 Exempel på orsaker till att studier exkluderas.

- Ej relevant studiedesign
- Bakgrundsmaterial
- Ej relevant population
- Ej relevant intervention
- Ej relevant kontrollintervention
- Ej relevant utfallsmått
- För kort uppföljningstid
- Dubbelpublikation
- Baslinjedata ofullständigt rapporterade

6. Kvalitetsgranskning av behandlingsstudier

VERSION 2012:1

Detta kapitel beskriver olika upplägg av behandlingsstudier, vilka metodologiska problem som är förknippade med respektive studietyp och hur man granskar studierna.

Randomiserade studier värderas generellt sett högst när undersökningen gäller behandlingseffekter. Anledningen är att de ger större förutsättningar för att kontrollera för faktorer som inte har med själva interventionen att göra. När det gäller att bedöma risker kan dock observationsstudier eller fall–kontrollstudier vara att föredra. Tvärsnittsstudier och fallserier utan kontroll har sämre tillförlitlighet och ingår sällan i SBU-projekt.

Avsnittet avslutas med metodproblem för systematiska översikter och hur översikterna granskas.

Granskningsmallar

Granskningen syftar till att bedöma i vilken utsträckning studiens resultat beror av systematiska fel (bias). SBU har utvecklat separata granskningsmallar (checklistor) för de olika studietyperna (Bilaga 2–3). Mallarna tar i ett antal frågor upp olika kvalitetsaspekter som påverkar studiernas tillförlitlighet. De är konstruerade så att det önskvärda svaret på varje fråga är ”ja”. I praktiken kan det ofta vara svårt att besvara frågorna. Information kan t ex vara otydlig. Det är därför viktigt att projektgruppen tränar gemensamt på mallarna. Det ger möjlighet att reda ut otydligheter och hur frågorna ska tolkas. Det kan vara värdefullt att göra kappa-mätningar vid några tillfällen under projektets lopp för att säkerställa att bedömningarna görs på likartat sätt.

Innan själva granskningen påbörjas bör projektgruppen besluta om några aspekter är speciellt viktiga för kvaliteten medan andra kanske inte ens är relevanta. Det är sällan som studier kommer att uppfylla samtliga kvalitetskriterier, utan bilden kommer att vara blandad. Mallen visar vilka brister en studie har – sedan måste granskarna avgöra i hur stor utsträckning bristerna påverkar tillförlitligheten.

Observera att granskningsmallarna *enbart är ett stöd* för att bedöma studiernas kvalitet. De är inte avsedda att användas så att man sätter krav på ett visst antal ”ja” för att studien ska åsättas en viss studiekvalitet.

Det kan vara klokt att spara den ifyllda granskningsmallen tillsammans med studien och att skriva ner de överväganden som gjorts i samband med granskningen. Detta underlättar vid den kommande skrivningen av rapporten.

För att bedöma risken för systematiska fel i enskilda studier används del A (A1–A6) i granskningsmallen. För att resultaten även ska kunna användas i en sammanvägd bedömning enligt GRADE krävs en del ytterligare information i form av sammanställningar av samtliga ingående studier. Det gäller bristande överensstämmelse mellan studierna (B), studiens överförbarhet (C) samt granskning av studiernas precision (D), publikationsbias (E), effektstorlek (F), dos–responssamband (G) och sannolikhet att effekten är underskattad (H). Sammanvägningen enligt GRADE sker vid ett senare tillfälle än vid granskningen av en enskild studie, men det kan vara lämpligt att vid läsningen av en enskild studie samtidigt kommentera dessa faktorer (B–H).

Randomiserad kontrollerad studie (RCT)

Randomisering innebär att man slumpmässigt fördelar studiedeltagarna mellan den experimentella behandlingen (försöksgruppen) och kontrollbehandlingen. Fördelen med randomisering är att slumpförfarandet ska leda till att alla andra faktorer än behandlingen ska fördelas lika på de båda grupperna. Skillnaden i effekt mellan försöks- och kontrollgrupperna är då sannolikt bara beroende på behandlingen (resonemanget gäller även för studier med mer än två olika grupper).

Det finns en internationell överenskommelse, CONSORT statement, om hur randomiserade studier bör redovisas [1]. CONSORT kan även ses som stöd för vilka aspekter av en studie som är viktiga att bedöma och är ett underlag för SBU:s granskningsmall. Mallen finns i Bilaga 2.

Randomisering

Randomiseringsförfarandet påverkar mycket hur tillförlitlig studien är. En bra randomisering är utformad så att fördelningen av försökspersoner till de olika grupperna inte kan påverkas. Proceduren för randomiseringen måste därför vara detaljerat beskriven. Speciellt i äldre studier (publicerade före år 1990) är det vanligt att randomiseringen inte beskrivs alls. Här får man avgöra från fall till fall hur den bristande informationen påverkar studiekvaliteten. En näraliggande aspekt till randomisering är gruppernas jämförbarhet, dvs om det fanns några kända skillnader mellan grupperna vid ”baseline”, trots randomiseringen (Exempel 6.1). Riskerna för selektionsbias tas upp i granskningsmallen (Bilaga 2, A1 Selektionsbias).

Behandlingsbias och blindning

Den andra faktorn som har stor betydelse för studiens kvalitet är graden av blindning. Patientens eller läkarens förväntningar på resultatet av behandlingen kan påverka utfallet eller mätningen av utfallet. Det gäller särskilt för mjuka variabler som t ex livskvalitet eller patientens skattning av symtombörda. Risken för påverkan på effektmått som död eller

Exempel 6.1 En äldre kontrollgrupp.

I en studie där man prövade ifall behandling med alfablockerare alfuzosin ökade chansen för framgångsrik dragning av kateter råkade kontrollgruppen bli i genomsnitt fem år äldre jämfört med behandlingsgruppen [2]. Eftersom ålder visades vara den starkaste prediktorn för misslyckad kateterdragning oavsett behandling så blev resultatet svårbedömt.

fraktur är lägre. Så många aktörer som möjligt i studien bör därför vara blindade, dvs inte känna till vilken behandling som ges (experimentell eller kontroll). Det ideala är om alla parter (vårdgivare, patient, den som mäter effekten och den som redovisar resultaten) är blindade, så kallad trippelblind studie.

I praktiken kan det vara svårt att blinda studien. Det gäller t ex inom kirurgin. Om det är etiskt försvarbart kan man använda så kallad ”sham”-kirurgi (”låtsaskirurgi”) för att minimera systematiska fel pga patienternas förväntningar. Andra exempel är livsstilsinterventioner, psykoterapi och fysioterapi. Även om det inte går att blinda behandlare och patient går det att blinda dem som registrerar och utvärderar resultaten. Denna grad av blindning kan användas som ett minimikrav.

Frågor kring behandlingsbias, blindning och följsamhet återfinns i granskningsmallen (Bilaga 2, A2 Behandlingsbias).

Bedömning av resultatredovisningen

Inför bedömningen av hur tillförlitliga själva resultaten är (Bilaga 2, A3 Bedömningsbias och A5 Rapporteringsbias) behöver vi tillgång till uppgifter om grundförutsättningarna för studien. Vilket är det primära effektmåttet? Hur stor effekt förväntas behandlingen ge?

Det är sedan viktigt att de resultat som författarna lyfter fram baseras just på det primära effektmåttet. Om det inte finns någon statistiskt säkerställd skillnad i det primära effektmåttet är det inte ovanligt att resultaten avser ett annat effektmått där författarna sett en skillnad. Andra sätt att försöka trolla bort ett negativt resultat är att utföra post hoc subgruppsanalyser och blåsa upp eventuella skillnader mellan interventions- och kontrollarm i subgrupper.

Inför denna del av granskningen bör granskarna ha definierat vad som är minsta kliniskt relevanta effekt. Stora studier som påvisar en statistiskt signifikant men icke kliniskt relevant effekt får anses ha ett mindre värde.

Bortfall

Den tredje viktiga aspekten på studiekvalitet är bortfall (Bilaga 2, A4 Bortfallsbias). Tillförlitligheten är beroende av om de som ingår i studien följs upp under hela tiden och kan ingå i analysen. Ett stort bortfall är bl a problematiskt om effektmåttet är symtombaserat. Det kan tänkas att patienter utan symtomlindring avbryter studien i större utsträckning än de som förbättrats. Särskilt allvarligt är det om bortfallet skiljer sig mellan experiment- och kontrollgruppen. För läkemedelsstudier har SBU som riktmärke tillämpat följande ungefärliga gränser:

- <10 procent: bortfallet påverkar knappast tillförlitligheten och därmed inte heller studiekvaliteten
- >30 procent: bortfallet påverkar tillförlitligheten så mycket att studien saknar informationsvärde. Studien exkluderas.

Projektgruppen bör definiera på förhand hur stora bortfall som kan accepteras och i vilken utsträckning de påverkar studiens kvalitet. I vissa fall kan bedömningen påverkas positivt om författarna med hjälp av t ex en rimlig bortfallsanalys kan argumentera för att bortfallet inte stört resultatet.

I studier där man använder kontinuerliga variabler eller skalor använder man sig i vissa fall av beräkningsmetoden *LOCF* (”last observation carried forward”) för att kompensera för bortfall. Det senast uppmätta resultatet antas då gälla även för de senare tidpunkterna när data saknas. Det finns även andra statistiska metoder för att korrigera för bortfall. Känslighetsanalyser är sannolikt det mest användbara eftersom det ger en fingervisning om effekten kvarstår även under värsta tänkbara betingelser. Ett sådant är t ex att ingen i bortfallet förbättrades. Kvarstod effekten även i det värsta scenariet?

Intressekonflikter

Studierna måste också granskas utifrån en analys av om eventuella ekonomiska eller andra intressen kan påverka risken för att resultaten inte är tillförlitliga (Bilaga 2, A6 Intressekonflikter). Sådana intressekonflikter gäller kanske framför allt industrisponsrade metoder (läkemedel, medicinteknik) men även när forskare studerar en metod som de själva utvecklat.

Underlag för sammanvägd bedöming enligt GRADE

I en samlad bedömning (Kapitel 10) granskas bristande överensstämmelse mellan studierna (B), studiens överförbarhet (C) samt granskning av studiernas precision (D), publikationsbias (E), effektstorlek (F), dos–responssamband (G) och sannolikhet att effekten är underskattad (H). Några kommentarer kring dessa bedömningar kan vara värda att

göra. Granskningen av studiens överförbarhet ska t ex fokusera på om studierna är tillräckligt lik den population som SBU/HTA-rapporten behandlar.

Precisionen i data beror i huvudsak av studiens storlek, insjuknandefrekvensen och effektstorleken. Små studier kan vara problematiska av flera skäl. Dels har de svårare att visa statistiskt signifikanta resultat, dels finns det risk för att de är mindre välplanerade än större. Risken för obalanser i kända *och* okända bakgrundsfaktorer ökar. För små studier finns också en stor risk för typ 2-fel (nollhypotesen accepteras fast den är falsk). Det innebär att författarna inte lyckas påvisa en sann behandlingseffekt eftersom studiepopulationen är för liten för att åstadkomma en statistiskt säkerställd effekt.

Det är viktigt att författarna angett ”tidpunkter” för slutanalys och eventuella interimsanalyser och hur dessa hanterats statistiskt. Annars kan man misstänka att författarna har adderat deltagare successivt till studien tills de har fått en statistisk signifikans.

Själva uträkningen av effekten kan i princip göras på två sätt. I en *per protokollanalys* beräknas resultaten enbart för de personer som har följt hela studieprotokollet (”completers”). *ITT-analys* (”intention to treat”) är en mer konservativ beräkningsmetod som syftar till att minska risken för överskattning av behandlingsresultatet. ITT innebär att alla personer som har randomiserats följs upp inom sin behandlingsgrupp oavsett om de har fått rätt behandling eller inte. ITT är den bästa metoden för att mäta effekter av en intervention. Däremot är ofta per protokollanalys att föredra för att mäta biverkningar, eftersom man då bara mäter biverkningarna för dem som har fått behandlingen eller exponerats för en riskfaktor. Utspädningseffekter kan annars göra att man missar eventuella risker. I ”non-inferiority”-studier ska såväl per protokoll-analysen som ITT-analysen redovisas.

Effektstorlek (F) och eventuella dos–responssamband (G) påverkar naturligtvis tilltron till resultaten.

Observationsstudier

Observationsstudier kan planeras och utföras på olika sätt och metodologiskt brukar man dela in dem i kohortstudier, fall–kontrollstudier och tvärsnittsstudier. Kohortstudier som inkluderas i SBU:s granskningar ska vara kontrollerade, dvs ha en jämförelsegrupp.

Riktlinjer och råd för hur man bör värdera kohort- och andra observationsstudier har utvecklats internationellt, t ex STROBE [3]. Den granskningsmall som SBU använder för observationsstudier finns i Bilaga 3.

När det gäller observationsstudier är kvalitetsproblemen oftast allvarligare för små studier, de med historiska kontroller och studier som inte justerat för viktiga förväxlingsfaktorer, så kallade ”confounders”. Selektionsproblemen kan vara särskilt stora när det gäller preventiva och icke-akuta åtgärder där välinformerade patienter kan efterfråga specifika insatser.

6

Kohortstudier

Kontrollerade kohortstudier jämför en grupp som fått behandling eller utsatts för en risk med en grupp som fått alternativ eller ingen behandling respektive inte utsatts för en risk. Kohort betyder grupp och i en kohortstudie följer man individer över tiden framåt (prospektivt) för att se hur det går för dem. När studien startar insamlas oftast en mängd uppgifter om individerna i respektive grupp. Ålder, kön, socioekonomisk situation, levnadsvanor samt sjukdomar är t ex viktiga och grundläggande uppgifter. Sambandet mellan rökning och flera allvarliga sjukdomar påvisades t ex redan på 1950-talet genom att en kohortstudie följde hur det gick för brittiska läkare som rökte respektive inte rökte [4]. Med tanke på hur rökvanorna sedan minskat som en följd av denna kunskap är det sannolikt den enskilda studien som räddat flest liv i världen. Våra kunskaper om riskfaktorer för hjärt- och kärlsjukdomar som högt blodtryck, höga kolesterolvärden, rökning och fetma kommer från flera stora kohortstudier som t ex Framinghamstudien i USA och ”1913 års män i Göteborg”.

En nackdel med kohortstudier är att de kan bli kostsamma och svåra att genomföra vid undersökningar av sällsynta sjukdomar eller då det tar lång tid innan man kan mäta utfallet. Det krävs då stora studiepopulationer och långa uppföljningstider. Fall–kontrollstudier kan då vara en mer lämplig och kostnadseffektiv metodik (se nedan). I Sverige och framför allt övriga nordiska länder med väl utvecklade hälsodata- och kvalitetsregister finns dock ofta data redan insamlade för stora studiepopulationer. Det gör att kohortstudier många gånger även kan användas vid studier av sällsynta sjukdomar och krav på långa uppföljningstider (Exempel 6.2).

Exempel 6.2 Registerbaserad kohortstudie.

En dansk registerbaserad kohortstudie analyserade risken för autism bland barn som vaccinerats för mässling, påssjuka och röda hund. Totalt följdes mer än 537 000 barn under åtta års tid. Ingen ökad risk för autism kunde visas för de barn som vaccinerats jämfört med de ovaccinerade [5].

Fall–kontrollstudier

Fall–kontrollstudier är tillbakablickande (retrospektiva) till sin karaktär. Här letar man först upp fallen, dvs individer som drabbats av utfallet (t ex sjukdom eller död). Därefter jämförs de med individer som inte drabbats av detta utfall, dvs kontrollgruppen. Man kan sedan studera om grupperna skiljer sig åt avseende riskfaktorer eller vilken behandling de fått. Fall och kontroller måste representera samma studiebas och urvalet av kontroller måste ske strikt oberoende av eventuell exponering för behandlingen.

Vanligen måste man samla in data retrospektivt. Datainsamlingen kan göras med hjälp av intervjuer av fall och kontroller eller med hjälp av patientjournaler eller registerdata (Exempel 6.3). Problemet med intervjuer är att man inte alltid kommer ihåg allt som hänt tidigare i livet och att man tenderar att komma ihåg eller återge minnen olika beroende på om man tillhör patient- eller kontrollgrupp.

Exempel 6.3 Fall–kontrollstudie.

En svensk fall–kontrollstudie avsåg frågan om aspirin och NSAID-preparat kan reducera risken för magcancer [6]. Man interjuade 567 personer med magcancer och 1 165 personer utan cancer (kontroller) om bl a deras användning av smärtstillande medel. De som använde aspirin hade en minskad risk för magcancer (OR=0,7, oddskvot) efter kontroll för kön, ålder och socioekonomisk status. Risken minskade med ökad användning av aspirin. Däremot sågs inget samband mellan magcancer och andra smärtstillande medel.

Valet mellan fall–kontroll- eller kohortstudie betingas i stor utsträckning av praktiska och ekonomiska aspekter. Kohortstudier kan ofta belysa många olika hypoteser om en adekvat datainsamling skedde vid studiestart. Vid ovanliga utfall är fall–kontrollstudier ofta effektivare än andra studieupplägg, men vid ovanliga exponeringar fungerar de sämre för att påvisa samband.

Tvärsnittsundersökningar

Tvärsnittsundersökningar innebär att man mäter förhållanden vid en tidpunkt eller ett tillfälle. De är ett bra sätt för att få en uppfattning av prevalensen av olika sjukdomar och om undersökningarna visar samvariation mellan olika exponering och sjukdomar. Statistiska centralbyråns (SCB:s) årliga undersökningar om levnadsförhållanden är ett bra exempel på en tvärsnittsundersökning. Där intervjuas ett slumpmässigt urval av svenska folket årligen om en mängd förhållanden. Data har t ex använts för att uppskatta prevalensen av sjukdomar som inte alltid är lätt att uppskatta utifrån registerdata.

Tvärsnittsundersökningar exkluderas ofta i SBU-projekt eftersom det oftast är mycket svårt att säga något om tids- och orsakssambanden mellan intervention/exponering och utfall.

Kvalitetsvärdering av observationsstudier (se Bilaga 3 för granskningsmall)

6

Det generellt svåraste metodproblemet att hantera vid observationsstudier jämfört med randomiserade kontrollerade studier är att man måste kontrollera för *alla* potentiellt viktiga faktorer som kan påverka utfallet. Grupperna i observationsstudier kan skilja sig systematiskt från varandra i viktiga avseenden som kan påverka utfallet (selektionsbias). Om man tillförlitligt mätt dessa skillnader på individnivå går det att med statistiska metoder kontrollera för dem. Om de inte har mätts går det inte att utesluta att eventuella effekter som observerats beror på bakomliggande skillnader mellan grupperna snarare än den behandlingsinsats man vill studera.

Selektionsbias

Såväl behandlande läkare som patient kan påverka vilka som hamnar i den exponerade (försöksgruppen) respektive oexponerade gruppen (kontrollgruppen). Läkaren kan välja ut patienter som bedöms ha störst nytta av åtgärden. Dessa kan vara såväl friskare som sjukare än de som inte får behandlingen. Välinformerade patienter som ofta är friskare kan ha större kunskap om olika behandlingsalternativ och kan därför kräva specifika insatser som andra patienter inte känner till. Kvalitetsgranskningen av observationsstudier måste därför ha fokus på att bedöma om grupperna är jämförbara, eller om det finns förväxlingsfaktorer (”confounders”, se nedan) eller andra systematiska fel (Bilaga 3, A1 Selektionsbias). Det är ofta särskilt viktigt att justera analyserna med hänsyn till patienternas hälsotillstånd och socioekonomiska status [7].

Man kan aldrig helt utesluta risken för selektionsbias vid observationsstudier, men när det gäller okända eller oväntade biverkningar så är det dock mindre sannolikt [8]. Det kan bl a bero på att läkaren i sådana fall inte har kunskap om riskerna och därigenom inte samma möjlighet att selektera vilka som ska få respektive inte få läkemedlet sett ur ett riskperspektiv. En illustration av hur resultaten kan påverkas av selektionsbias finns i Exempel 6.4. En systematisk metaanalys kring biverkningar visade att det inte fanns någon skillnad i effektstorlek mellan randomiserade studier och observationsstudier [9]. Författarnas slutsats var att systematiska översikter av biverkningar inte ska begränsas till någon specifik studietyp.

Förväxlingsfaktorer

Förväxlingsfaktorer (”confounding factors”) kan vara ett stort problem i observationsstudier. Det innebär att man drar slutsatser om samband mellan en sjukdom och en bakgrundsfaktor medan sjukdomen i verkligheten förorsakas av en annan variabel som samvarierar med den studerade bakgrundsvariabeln.

Orsaksförväxling kan kontrolleras på olika sätt. Vid uppläggning av en studie kan man begränsa materialet så att förväxlingsproblem inte uppstår, genom att utesluta personer med en av de två samvarierande variablerna. Alternativa sätt att kontrollera för ”confounders” är att stratifiera och matcha materialet eller kontrollera med hjälp av olika typer av multivariata statistiska metoder. Stratifiering av materialet innebär att man delar upp materialet i undergrupper med olika exponering för att minska risken för förväxling. Exempel på sådana variabler kan vara kön, ålder och rökning. Matchning är ett annat sätt att minska risken för förväxling. Matchning innebär att man väljer personer som är lika med avseende på de förväxlingsfaktorer man vill kontrollera. Med hjälp av statistiska metoder, t ex regressionsmodeller, kan man samtidigt konstanthålla för flera variabler som kan förväxlas. En annan metodik för att kontrollera för ”confounders” är ”propensity score” [10].

Metodologiska skäl talar för att sambanden i observationsstudierna snarast är underskattade. Om felklassificeringen i exposition, t ex blodtrycksnivåer, är oberoende av utfallet underskattas alltid sambandets styrka, så kallad ”regression dilution bias” [11]. Ett förhållande som t ex alltid gäller för kohortstudier där exponeringsinformation insamlas före observationsperioden. Sannolikt klassificeras även data mer noggrant i randomiserade kontrollerade studier än i observationsstudier.

Exempel 6.4 Risk för selektionsbias och behov av kontroll för förväxlingsfaktorer.

Risken för selektionsbias vid observationsstudier kan illustreras med det ofta citerade exemplet om östrogenbehandling och risk för hjärt- och kärlsjukdom. Flera observationsstudier hade pekat på att östrogenbehandling minskade risken för hjärt- och kärlsjukdom. Detta trots att man i en del studier försökt justera resultaten för en del riskfaktorer. En stor randomiserad kontrollerad studie visade att så inte var fallet. När man i observationsstudierna däremot justerade riskerna med hänsyn till skillnader i socioekonomi så försvann riskminskningen [12]. Detta var inte förvånande mot bakgrund av att flera andra studier visat att välutbildade kvinnor i större utsträckning än andra fick östrogenbehandling. Ifall de publicerade observationsstudierna hade haft med socioekonomiska förhållanden i sin analys hade man sannolikt inte dragit felaktiga slutsatser. Förhoppningen om att vitaminer kunde skydda mot hjärt- och kärlsjukdomar och lungcancer kunde inte heller påvisas när man i observationsstudier justerade för socioekonomi [13].

Biverkningsstudier

Vid utvärdering av olika åtgärder har observationsstudier tveklöst ett stort värde för att påvisa negativa effekter och risker, dvs biverkningar. Biverkningar är ofta till sin natur oväntade och en del allvarliga biverkningar är sällsynta, men av sådan dignitet att även mycket låga incidenstal är oacceptabla (Exempel 6.5). Randomiserade kontrollerade studier är sällan dimensionerade, vare sig i termer av urvalsstorlek, uppföljningstid, eller rapporteringsrutiner, för att fånga upp sådana biverkningar. Registerbaserade kohort- och fall–kontrollstudier med stora populationer är då ofta ett bra alternativ.

Exempel 6.5 Värdet av observationsstudier vid analys av biverkningar/risker.

Läkemedlet aprotinin har sedan år 1993 använts runt om i världen för att minska blödning vid bl a by-pass-operationer. Många små randomiserade kontrollerade studier hade inte sett några risker vid användningen av aprotinin. Däremot kunde stora kohort- och fall–kontrollstudier visa på en överrisk för njursvikt och död [14,15]. Det var dock först när en större randomiserad kontrollerad studie senare avbröts pga ökad dödlighet inom 30 dagar som medlet drogs in. Sannolikt hade ett antal dödsfall kunnat undvikas om man tagit dessa observationsstudier på större allvar och mer kritiskt granskat metaanalyser av små RCT [16]. Europeiska läkemedelsverket (EMA) har dock tillåtit användningen av aprotinin på särskilda indikationer.

Faktaruta 6.1 Läkemedelsverkets bedömning av biverkningsrapporter.

Läkemedelsverket gör en bedömning av biverkningsrapporter och graderar det kausala sambandet för de rapporterade biverkningarna som (a) säkert eller sannolikt; (b) möjligt; (c) osannolikt; eller (d) ej bedömbart, baserat på biologisk rimlighet, möjliga verkningsmekanismer, tidssamband och – mer sällan förekommande – resultatet av provokationsstudier där försökspersoner utsätts för en viss exponering eller inte. Vid en sådan bedömning görs inga försök att värdera de enskilda studiernas kvalitet eller den sammantagna evidensgraden.

Tabellera studierna

En huvuduppgift i projekten är att extrahera data ur studier med hög och medelhög studiekvalitet och sammanställa dem i tabeller. Syftet är att läsare av rapporten på ett enkelt sätt ska kunna få en överblick över inkluderade studier och hur de har bedömts. Ett annat syfte är att det underlättar det fortsatta arbetet genom att data blir strukturerade.

Om det vetenskapliga underlaget enbart består av studier med låg studiekvalitet ska dessa tabelleras.

Tabellerna ska ge information om referens, frågeställning, metod, urval, genomförande, resultat och metodologisk kvalitet. Tabellerna skrivs på engelska. Skälen är att underlätta för andra länder att tillgodogöra sig en del av SBU:s grundläggande arbete, att artiklarna oftast är på engelska och att man då inte behöver översätta de exakta uttrycken som används i artiklarna samt att tabeller huvudsakligen läses av vetenskapligt skolade personer med god kunskap i att läsa vetenskaplig litteratur. Eventuella tabeller i sammanfattningen ska dock alltid vara på svenska. I Tabell 6.1 ges exempel på hur en tabell kan konstrueras så att den innehåller relevant information.

Tabell 6.1 Exempel på tabellstruktur.

Author Year Reference Country	**Study design**	**Population, patient characteristics**	**Intervention**	**Follow-up period Drop out rate**	**Results**	**Study quality and relevance Comments**
	(RCT, CT, cohort, case control etc)	Inclusion/ exclusion criteria Setting No at baseline Male/ female	Intervention (I) (dose, interval, duration) Control (C) (active, placebo, usual care, etc)	(From baseline to follow-up, or from end of intervention to follow-up) Drop out (%)	Results (I, C) (Absolute difference, HR, RR, OR, p-value, confidence interval for the difference, sensitivity, specificity, observer reliability, cost-effectiveness, etc)	High, moderate or low study quality if appropriate

C = Control; CT = Controlled trial; HR = Hazard ratio; I = Intervention; OR = Odds ratio; RCT = Randomised controlled trial; RR = Risk ratio

Tabellen kan användas för att göra en kvalitativ bedömning rörande heterogenitet eller stora variationer mellan studierna. De ger också en samlad bild av kunskapsläget för en specifik frågeställning.

Systematiska litteraturöversikter och HTA-rapporter

Principiellt kan en forskningsfråga i projektet besvaras med en redan publicerad systematisk litteraturöversikt. Förutsättningarna är att:

- frågan överensstämmer helt med projektets frågeställningar
- litteraturöversikten är tillförlitlig.

I och med att projektgruppen (och därmed SBU) i praktiken ställer sig bakom slutsatserna i en annan systematisk översikt måste översikten underkastas en noggrann granskning.

Ofta har systematiska översikter sådana metodologiska brister att de inte kan användas, annat än som referenslista. I en tvärsnittsundersökning av kvaliteten på systematiska översikter som publicerades vintern 2007 saknades t ex kvalitetsgranskning i cirka 30 procent av översikterna [17].

I syfte att förbättra kvaliteten i systematiska översikter och rapporteringen av metaanalyser har en internationell grupp publicerat en rapport, PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses, www.prisma-statement.org), som är en vidareutveckling av det tidigare så kallade QUOROM statement. I PRISMA och därtill hörande dokument finns rekommendationer för hur tydligheten och transparensen kan förbättras. På sikt kommer därför sannolikt kvaliteten på systematiska översikter att förbättras.

Det första steget i bedömningen är att avgöra om den systematiska översikten är relevant. Överensstämmer översiktens frågor samt inklusions- och exklusionskriterier med projektets fråga? Om översikten har annat syfte eller andra kriterier exkluderas den.

Därefter bedöms översiktens vetenskapliga kvalitet. Den systematiska litteraturöversikten bör bl a innehålla:

- klart formulerade frågor
- en tydligt beskriven metod för litteratursökning och urval av artiklar
- kvalitetsbedömning av studier som uppfyller översiktens inklusions- och exklusionskriterier
- tabeller som redovisar data för inkluderade studier
- en sammanvägning av studiernas resultat med lämpliga metoder, exempelvis metaanalys
- formuleringar som tyder på att författarna har tagit hänsyn till de ingående studiernas vetenskapliga kvalitet i slutsatserna.

Även systematiska litteraturöversikter och HTA-rapporter granskas och värderas med hjälp av en granskningsmall (Bilaga 6). Granskningsmallen baseras på en internationellt utvecklad granskningsmall, AMSTAR [18,19]. AMSTAR består liksom övriga checklistor av en rad frågor som kan besvaras med ”ja”, ”nej”, ”kan inte svara” och ”ej tillämpligt”. Projektgruppen bör i förväg besluta vilka krav som måste besvaras med ”ja” (eller ”ej tillämpligt”) för att kvaliteten på översikten ska anses vara godkänd.

En viktig fråga att ta ställning till är om översikten fångat in alla relevanta artiklar. Om någon studie saknas kan det antingen bero på att sökstrategin inte är tillräckligt bra eller på publikationsbias. Intressekonflikter kan påverka vilka studier som inkluderats, något som kan vara svårt att kontrollera.

För ytterligare kvalitetssäkring rekommenderas att projektgruppen kontrollerar fakta och tolkningar genom att även läsa några av de studier som inkluderats i översikten. Anledningen är att det förekommer att studier har feltolkats vilket påverkar hela analysen. Studier som kodats som randomiserade kan vid närmare betraktande vara observationsstudier.

Tabell 6.2 sammanfattar vad som krävs för att SBU ska acceptera en systematisk översikt för att besvara en fråga.

Tabell 6.2 Bedömning av användbarhet av systematiska litteraturöversikter och HTA-rapporter.

Alternativ	Kvalitetsgranskning (enligt Bilaga 6)	Slutsats och evidensgradering	Åtgärd
Frågor och inklusions- och exklusionskriterier överensstämmer med projektets	Godkänd	Slutsatserna kan accepteras om inte senare tillkommen litteratur motsäger detta	Översikten inkluderas och kompletteras med eventuellt senare publicerade artiklar. Översikten evidensgraderas med GRADE
	Inte godkänd	Slutsatserna accepteras inte	Översikten exkluderas. Referenslista och annan information utnyttjas i det egna arbetet
Frågorna överensstämmer, men inte inklusions- och exklusionskriterier		Slutsatserna accepteras inte	Översikten exkluderas. Referenslista och annan information utnyttjas i det egna arbetet

Referenser

1. Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMC Med 2010;8:18. http://www.biomedcentral.com/1741-7015/8/18

2. McNeill SA, Daruwala PD, Mitchell ID, Shearer MG, Hargreave TB. Sustained-release alfuzosin and trial without catheter after acute urinary retention: a prospective placebo-controlled. BJU Int 1999;84:622-7.

3. von Elm E, Egger M, Altman DG, Pocock SJ, Vandenbroucke JP. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007;335:806-8.

4. Doll R, Hill AB. The mortality of doctors in relation to their smoking habits: a preliminary report. BMJ 1954;1:1451-5.

5. Madsen KM, Hviid A, Vestergaard M, Schendel D, Wohlfahrt J, Thorsen P, et al. A population-based study of measles, mumps, and rubella vaccination and autism. N Engl J Med 2002;347:1477-82.

6. Akre K, Ekström AM, Signorello LB, Hansson L-E, Nyrén O. Aspirin and risk for gastric cancer: a population-based case-control study in Sweden. Br J Cancer 2001; 84:965-8.

7. Rosén M, Axelsson S, Lindblom J. Släng inte ut observationsstudier med badvattnet. Bedöm deras kvalitet istället. Läkartidningen 2008;105:3191-4.

8. Vandenbroucke JP. When are observational studies as credible as randomised trials? Lancet 2004;363:1728-31.

9. Golder S, Loke YK, Bland M. Meta-analyses of adverse effects data derived from randomized controlled trials as compared to observational studies: methodological overview. PLoS Medicine 2011;8:1-13.

10. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998;17:2265-81.

11. MacMahon S, Peto R, Cutler J, Collins R, Sorlie P, Neaton J, et al. Blood pressure, stroke and coronary heart disease. Part 1, Prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. Lancet 1990;335:765-74.

12. Humphrey LL, Chan BK, Sox HC. Postmenopausal hormone replacement therapy and the primary prevention of cardiovascular disease. Ann Intern Med 2002;137:273-84.

13. Lawlor DA, Davey Smith G, Brucksdorfer KR, Kundu D, Ebrahim S. Those confounded vitamins: what can we learn from the differences between observational versus randomised trial evidence? Lancet 2004;363:1724-7.

14. Mangano DT, Tudor JC, Dietzel C. The risk associated with aprotonin in cardiac surgery. N Engl J Med 2006;354:353-65.

15. Schneeweiss S, Seeger JD, Landon J, Walker AM. Aprotinin during coronary-artery bypass grafting and risk of death. N Engl J Med 2008;358:771-83.

16. Rosén M. The aprotinin saga and the risks of conducting meta-analysis on small randomised controlled trials – a critique of a Cochrane review. BMC Health Serv Res 2009;9:34.

17. Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. PLoS Med. 2007;4:e78.

18. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol 2007;7:10.

19. Shea BJ, Hamel C, Wells GA, Bouter LM, Kristjansson E, Grimshaw J, et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. J Clin Epidemiol 2009;62:1013-20.

7. Kvalitetsgranskning av diagnostiska studier

VERSION 2012:1.1

Bakgrund

Diagnostiska test syftar till att identifiera personer med sjukdom eller hälsoproblem. Det räcker dock inte att tillförlitligt diagnostisera en individ. Om det inte finns effektiva behandlingsmetoder är nyttan mycket begränsad eller kan t o m vara negativ.

Screening är en form av diagnostik som kan utvärderas på i stort samma sätt som diagnostik vid misstänkt sjukdom. Den väsentliga skillnaden är att populationsbaserad screening genomförs där en specifik befolkningsgrupp inbjuds och där prevalensen vanligtvis är mycket lägre än för personer som sökt vård för ett misstänkt problem. Vid screening är därför möjligheterna att förutsäga vilka som är sjuka respektive friska sämre och problemen med falskt positiva större. WHO har i en gammal, men fortfarande aktuell, rapport angett vilka kriterier som ska vara uppfyllda för att screening ska kunna introduceras [1]. Dit hör t ex att det ska finnas behandlingar som påverkar prognosen, och att nyttan ska vara större än uppskattade risker och kostnader.

Tester (inkluderande kliniska fynd, symtom, bilder och biokemiska tester) kan användas för flera olika ändamål, t ex för att bedöma risk, prognos, följa ett sjukdomsförlopp eller värdera effekten av behandling. I detta kapitel fokuseras på diagnostiska test.

Det principiella tillvägagångssättet för en systematisk översikt om diagnostik är detsamma som för en översikt om behandling, dvs:

- definition av syften, formulering av frågeställning(ar)
- sökning och selektion av studier
- granskning och kvalitetsbedömning/gradering
- dataextraktion, tabellering
- statistisk analys, sammanställning
- tolkning och presentation av resultat.

Det finns dock flera skillnader mellan en översikt om diagnostik och en översikt avseende interventionsstudier:

- I behandlingsstudier är utfallet t ex överlevnad eller sjuklighet. I diagnostiska studier är utfallet istället intermediärt, dvs det har inget omedelbart, direkt värde för patienten.

- Litteratursökningen innebär ofta en större utmaning beroende på att indexering av relevanta studier inte är lika tydlig som för interventionsstudier. Man behöver därför ofta upprepa sin sökning med kompletterande söktermer.

- Kriterier för att bedöma studiekvalitet (potentiella bias) skiljer sig delvis från dem som används för interventionsstudier. Granskning av inkluderade studier görs med en speciell mall (QUADAS, se Bilaga 4) [2].

- I motsats till interventionsstudier, där randomiserad studiedesign är vanlig, finns ingen speciell dominerande design för att bestämma diagnostisk tillförlitlighet hos ett test. Randomiserad studiedesign förekommer, men ofta är studierna av tvärsnittskaraktär eller icke randomiserade observationsstudier (kohortstudier), vilket ställer stora krav på noggrann beskrivning av den studerade patientpopulationen.

- Rapporteringen är ofta bristfällig [3,4]. Det gäller såväl beskrivning av populationen, utförda tester och redovisning av resultat. Det saknas konsensus kring hur data ska sammanställas i en diagnostisk studie, vilket kan begränsa möjligheterna att göra metaanalyser.

- Diagnostiska studier uppvisar oftast heterogena resultat beroende på olika patientsammansättning, och på att använt tröskelvärde varierar. Att väga samman resultaten i metaanalyser är därför i flera fall inte lämpligt.

- Statistisk analys och sammanställning av resultat från enskilda studier skiljer sig från dem som används för interventionsstudier. Det beror på att resultaten består av par som är sinsemellan beroende (t ex sensitivitet och specificitet). Detta beskrivs närmare i nästa avsnitt.

Termer och mått

Diagnostisk tillförlitlighet (”accuracy”)

De mått, metoder och presentationssätt som används för att värdera diagnostiska test redovisas mer utförligt i Bilaga 9.

Diagnostisk tillförlitlighet avser ett tests förmåga att skilja på individer med sjukdom (eller mer generellt ett sökt tillstånd) och dem utan sjukdom. Det test som ska utvärderas kallas *indextest*. Ibland jämförs ett nytt test med ett test som utgör standard i klinisk praxis (jämförande indextest). De flesta studier fokuserar på att undersöka tillförlitligheten hos ett isolerat test, men i praktiken existerar aldrig ett test i ett vakuum utan

ingår som en del i en diagnostisk process. Testets roll då multipla tester används belyses i Faktaruta 7.1.

Faktaruta 7.1 Multipla tester.

I praktiken används ofta flera tester för att bedöma sannolikheten för sjukdom hos en patient som söker vård pga symtom. Det kan vara praktiskt att tänka igenom vilken funktion det undersökta testet har i den diagnostiska kedjan. Används det ensamt eller tillsammans med andra tester? Multipla tester kan i princip användas på två sätt:

- parallell testning (dvs alla tester på en gång) där ett positivt testresultat för *något av testerna* indikerar sjukdom.
- seriell testning (konsekutiv) testning, där beslutet att gå vidare med nästa test beror på utfallet av tidigare test(er). Här måste *alla tester* ge positivt testresultat för att diagnosen ska kunna ställas, eftersom den diagnostiska processen stoppas vid ett negativt resultat, se Figur 7.1.1.

Strategi	Händelseföljd	Konsekvenser
Parallell testning	Test A *eller* Test B *eller* Test C är positivt: A → +, − B → +, − C → +, −	Sensitivitet ↑ Specificitet ↓
Seriell testning	Test A *och* Test B *och* Test C är positiva: A → + B → + C → + − − −	Sensitivitet ↓ Specificitet ↑

Figur 7.1.1 Parallell och seriell testning. Parallell testning med multipla tester ökar i allmänhet sensitiviteten, medan specificiteten sjunker, och andelen falskt positiva testresultat ökar. Seriell testning maximerar specificitet, medan sensitiviteten sjunker, och andelen falskt negativa testresultat ökar [5].

I praktiken beställs ofta flera tester samtidigt, speciellt när snabb bedömning är nödvändig, t ex för patienter som lagts in på sjukhus eller vid akuta tillfällen, och det är inte alltid lätt att avgöra hur den existerande diagnostiska kedjan ser ut, och vilken roll ett enskilt test har. Existerande diagnostik kan vara grundad på riktlinjer, men ibland finns ingen konsensus för den optimala ordningen efter vilken tester ska göras. Man kan då tvingas göra antaganden om var ett test är placerat i den diagnostiska kedjan.

Tillförlitligheten hos indextestet bedöms genom att jämföra det med ett *referenstest* eller *referensstandard* (ibland benämnt ”gold standard”). För de flesta, om inte alla, medicinska tillstånd saknas en tydlig och helt felfri guldstandard, och begreppet referensstandard är därför ett bättre uttryck. Referensstandarden ska representera det bästa tillgängliga sättet för att påvisa sjukdomen eller tillståndet ifråga. Referensstandarden är ofta bristfällig eller saknas. Val av relevant referensstandard är en kritisk del när man ska fastställa inklusionskriterier, se Faktaruta 7.2.

Faktaruta 7.2 Hur gör man när referensstandarden är bristfällig eller saknas?

När det finns tydliga kriterier för det ”sanna” tillståndet (referensstandard eller referenstest), talar man om kriterievaliditet. Ett exempel är histologisk undersökning efter biopsi av bröstvävnad för att fastställa närvaro/frånvaro av bröstcancer. Ofta saknas dock en entydig eller en acceptabel referensstandard. Olika lösningar har föreslagits i situationer där referensstandarden är bristfällig eller saknas [6,7]. En referensstandard kan konstrueras (”construct validity”) genom att använda:

- **Sammansatt referensstandard.** Här kombineras flera (var för sig bristfälliga) referensstandard till ett sammansatt mått, som anses bättre diskriminera mellan sjukdom/ej sjukdom än enskilda referensstandard. För att bestämma närvaro/frånvaro av sjukdom används *fördefinierade regler*, där olika definitioner kan användas beroende på karakteristika hos de enskilda referenstesterna. Den enklaste modellen är att två referenstester appliceras på alla patienter. Oftast definierar man närvaro av sjukdom om något av referenstesterna är positivt. Ett exempel är en studie där man undersökte tillförlitligheten hos ett indextest för antigenanalys (”enzyme immunoessay analysis”, EIA) för att diagnostisera *Chlamydia trachomatis*-infektion [8]. Två referenstester kombinerades: odling och DNA-analys (”polymerase chain reaction”, PCR). EIA och de två referenstesterna applicerades på samtliga patienter. *Chlamydia trachomatis*-infektion ansågs föreligga om antingen odling eller PCR visade positivt resultat med EIA. Om båda referenstesterna var negativa, var diagnosen ej infektion.
- **Panel- eller konsensusdiagnos.** Här kombineras resultat av olika tester och andra resultat eller kliniska karakteristika och prognostisk information, som tillsammans ger en pragmatisk validering av sjukdomen. Validering av referensstandarden bygger då på en stor mängd empiriska data och bestäms ofta genom internationella konsensusprocedurer med *expertpaneler* eller med så kallad *Delfiprocedur* [7]. Ett exempel är DSM-IV – kriterier 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 [9].

Exemplet fortsätter på nästa sida

Faktaruta 7.2 Fortsättning.

- **Statistiska modeller.** Här kombineras klinisk information och andra testresultat i statistiska modeller, som genererar en sannolikhet för att t ex hjärtsvikt föreligger [7].

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.

Ibland kan man justera och korrigera för en bristfällig referensstandard. Det kräver att man har tillgång till god information om vad bristfälligheten består i, och hur stort felet är. En annan möjlighet är att göra en ”optimistisk” och en ”pessimistisk” beräkning av sensitivitet och specificitet.

Utfallsmått

De klassiska måtten för att beskriva tillförlitlighet hos diagnostiska test är *sensitivitet* och *specificitet*. Utifrån dessa och prevalens kan samtliga andra mått beräknas, t ex sannolikhetskvoter (”likelihood ratio”), diagnostisk oddskvot (DOR), samt positivt och negativt prediktionsvärde.

För att beräkna sensivitet och specificitet krävs binära (dikotoma) variabler. När ett diagnostiskt test består av kontinuerliga värden, t ex blodsockernivå, får ett visst gränsvärde bestämma vilka som ska definieras som diabetiker. Gränsen för vad som ska betecknas som sjukt eller friskt är dock inte självklar. Oftast väljs ett tröskelvärde, en ”cut-off”, där man bedömer att sannolikheten för att skilja mellan sjukdom/ej sjukdom är störst. Denna något godtyckliga gräns påverkar hur många sant och falskt positiva respektive negativa man kommer att få. Sänks gränsvärdet för blodsockernivån, kommer man att få fler falskt positiva och färre falskt negativa. Om gränsvärdet höjs blir resultatet det omvända, se Kapitel 9.

Det är väl känt att diagnostisk tillförlitlighet inte är en fix och stabil egenskap hos ett test i alla sammanhang. Sensitivitet och specificitet kan variera hos olika subgrupper av patienter, olika sjukdomsspektra eller olika kliniska situationer (primärvård, specialistvård, sjukhusvård). Den kan också variera beroende på olika tolkning av testet, och den kan vara beroende av föregående tester. Detta är viktigt att beakta då man beslutar vilka populationer som ska inkluderas i översikten enligt PICO (se avsnittet ”Formulera inklusions- och exklusionskriterier”) [3].

Hur ska tillförlitligheten hos ett test bedömas?

Vad som ska anses vara tillräckligt god tillförlitlighet hos ett diagnostiskt test beror på i vilket sammanhang testet används. Tester ger sällan en hundraprocentigt säker diagnos men kan ge tillräcklig information för att fastställa eller utesluta en diagnos på ett pragmatiskt sätt. En diagnos kan med andra ord vara tillräckligt säker för att den förväntade nyttan av att behandla patienten överväger de förväntade konsekvenserna av att inte behandla. Om både sensitivitet och specificitet hos ett test är högre än för ett annat test, så väljer man förstås det första. Men ofta tvingas man göra kompromisser och välja antingen hög sensitivitet eller hög specificitet. Då får man bedöma vilken typ av feldiagnos som får minst allvarliga konsekvenser. Exempel på detta ges i Tabell 7.1. Etiska aspekter, risk för biverkningar vid behandling och behandlingskostnader måste också vägas in vid bedömningen.

Tabell 7.1 Avvägning mellan krav på sensitivitet respektive specificitet.

Utfall	Riskbedömning
”Skador” då friska klassificeras som sjuka = krav på hög specificitet	Risker med att behandla friska individer Exempelvis vid fosterdiagnostik, eller där eventuell efterföljande behandling innebär stora risker, är kraven på specificitet mycket höga
”Skador” då sjuka klassificeras som friska = krav på hög sensitivitet	Individens risk för att inte bli behandlad, befolkningens risker (smittspridning), ärftliga risker Exempelvis för smittsamma sjukdomar bör sensitiviteten ofta vara hög

Analytisk noggrannhet (tillförlitlighet)

Det är viktigt att skilja mellan analytisk tillförlitlighet och diagnostisk tillförlitlighet. Den analytiska noggrannheten avser ett mätinstruments eller en mätmetods förmåga att ge användbar och pålitlig information om ett testresultat och bör ligga till grund för alla tester. Ett tillförlitligt mätinstrument eller mätmetod ska ha *god precision* och *inga systematiska fel.* Precisionen bestäms genom upprepade mätningar av samma patientprov och överensstämmelsen uttrycks ofta som variationskoefficienten (standarddeviationen/medelvärdet). Bristande reproducerbarhet av testresultat kan också bero på observatörsvariation vid tolkning av ett test. Det räcker dock inte med att precisionen är god. Ett mätinstrument bör också vara utan *systematiska fel.* Ett systematiskt fel kan bero på bristande kalibrering av mätinstrumentet. Ett exempel på detta är mätning av plasmakoncentrationen av kreatinin för att skatta njurfunktionen. Om analysmetoden för att bestämma kreatinin inte är kalibrerad mot en standard, är risken stor att det uppstår ett systematiskt fel. Storleken på det systematiska felet uttrycks ofta som bias, som i det här sammanhanget betyder medel- eller mediandifferensen mellan ett skattat värde (index-

metod) och ett ”sant” värde (referensmetod). Denna typ av frågeställning ingår sällan i SBU:s rapporter.

Definiera syften, formulera frågor

Som vid alla systematiska översikter är en genomtänkt och tydlig formulering av frågan/frågorna väsentlig. Det underlättar både sökning och selektion av för översikten relevanta studier.

Olika typer av frågor för diagnostiska studier

För läkemedel finns en strängt reglerad internationell standard (hierarkisk fyrfasmodell), 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). Något motsvarande finns inte för diagnostiska test, men åtskilliga motsvarande modeller för utvärdering av diagnostiska tester har föreslagits [10]. En av dem har en tydlig hierarkisk struktur, och består av fyra typer av frågor som är viktiga att beakta och överväga.

Steg 1-fråga. Skiljer sig testresultaten hos patienter med det sökta tillståndet från testresultaten från friska individer?

Steg 1-studier är ofta fall–kontrollstudier där en grupp patienter som har tillståndet ifråga jämförs med en grupp som inte har tillståndet. Resultaten presenteras ofta som korrelation eller skillnader i medelvärden mellan sjuka och friska. Ett positivt utfall i en Steg 1-studie gör det naturligt att gå vidare till nästa steg.

Steg 2-fråga. Är sannolikheten större att patienter med ett visst testresultat har tillståndet ifråga jämfört med patienter med andra testresultat?

Här ändras alltså inriktningen på tolkningen till diagnostik. Resultaten presenteras som sensitivitet och specificitet. Också här är studiedesignen ofta av fall–kontrollkaraktär.

Steg 3-fråga. Kan testresultat skilja ut individer med respektive utan sjukdom hos en grupp patienter där det är kliniskt rimligt att misstänka sjukdom?

Exempel 7.1 visar exempel på resultat från Steg 1-, Steg 2- och Steg 3-studier. Exemplen 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 diagnostiskt test i klinisk praxis.

Steg 4-fråga. Blir utfallet (resultat till följd av testutfallet) bättre för patienter som genomgår testet jämfört med liknande patienter som inte genomgår testet?

Denna fråga avser det egentliga värdet av ett test för patienten. Utfallet mäts i hälso-resultat till följd av beslut om eventuella ytterligare diagnostiska test och eventuell

behandlingsintervention. Ibland kan nyttan vara uppenbar som t ex vid korrekt diagnos vid livshotande tillstånd. Relativt ofta handlar det emellertid om tester för tidig upptäckt av asymtomatisk sjukdom, t ex 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).

Dessa fyra steg eller faser kallas den diagnostiska forskningens arkitektur [11]. Vilken eller vilka frågor som ska ingå i en översikt beror på det aktuella kunskapsläget.

Exempel 7.1 Exempel som illustrerar hur resultat varierar beroende på i vilket steg studien är gjord [11].

Steg 1-studie. På ett sjukhus mättes plasmakoncentrationen av BNP (”B-type natriuretic peptide”) precursor i ett icke systematiskt urval (”convenience sample”) av patienter med olika kombinationer av förhöjt blodtryck, hjärtkammarhypertrofi och systolisk dysfunktion hos vänster hjärtkammare och friska kontrollpatienter. Man fann stora skillnader i BNP-koncentration mellan grupperna.

BNP-koncentration	Patienter med tydlig systolisk dysfunktion	Normala kontroller
Koncentration (pg/ml) av BNP precursor medianvärde (variation)	493,5 (248,9–909,0)	129,4 (53,6–159,7)

Slutsatsen var att mätning av BNP precursor var ett värdefullt diagnostiskt hjälpmedel för att fastställa dysfunktion hos vänster hjärtkammare.

Steg 2-studie. På ett annat sjukhus testades samma metod på en grupp patienter med känd hjärt- och kärlsjukdom och varierande grad av hjärtkammardysfunktion och normala kontrollpatienter.

BNP-koncentration	Patienter med känd sjukdom	Normala kontroller
Hög	39	2
Normal	1	25

Testresultat (95 % konfidensintervall):
Sensitivitet = 98 % (87; 100); Specificitet = 92 % (77; 98)
Positivt prediktionsvärde = 95 % (84; 99)
Negativt prediktionsvärde = 96 % (81; 100)

Resultaten är extremt hoppingivande! Men är de alltför optimistiska? Resultaten baseras på patienter med etablerad sjukdom jämfört med friska individer.

Exemplet fortsätter på nästa sida

Exempel 7.1 Fortsättning.

Steg 3-studie. Är BNP-test användbart för att diagnostisera patienter med misstänkt dysfunktion hos vänster hjärtkammare (DVH)? Detta undersöktes hos en grupp patienter som remitterats pga misstänkt hjärtsvikt. Patienterna (n=126) genomgick oberoende, blindade mätningar av BNP och ekokardiografi.

BNP-koncentration	Patienter med DVH enligt ekokardiografi	Patienter med normala värden enligt ekokardiografi
Hög (>17,9 pg/ml)	35	57
Normal (<18 pg/ml)	5	29

Prevalens DVH (före test): 40/126=32 %.
Testresultat (95 % konfidensintervall):
Sensitivitet = 88 % (74; 94); Specificitet = 34 % (25; 44)
Positivt prediktionsvärde = 38 % (29; 48)
Negativt prediktionsvärde = 85 % (70; 94)

Mätning av BNP var inte tillnärmelsevis lika bra när testet användes i klinisk praxis, och författarna konkluderar att rutinmässig mätning av BNP sannolikt inte förbättrar diagnostiken av hjärtsvikt (dysfunktion hos vänster hjärtkammare).

Ett annat sätt att beskriva olika steg i den diagnostiska processen [12,13] är:

1. teknisk tillförlitlighet (analytisk precision och validitet)
2. diagnostisk tillförlitlighet (”accuracy”)
3. effekten på det kliniska beslutsfattandet
4. effekten av behandlingen på patientens välbefinnande
5. effekt på utfallet av behandlingen av patienten
6. kostnad–nytta, kostnadseffektivitet.

Denna modell kan vara användbar för att skilja mellan olika typer av studier, men den kan inte ses som en nödvändig sekvens för utvärderingar, eftersom utvärdering av tester sannolikt inte är linjär utan snarare cyklisk och repetitiv [10]. Separata systematiska översikter kan göras för var och en av dessa steg.

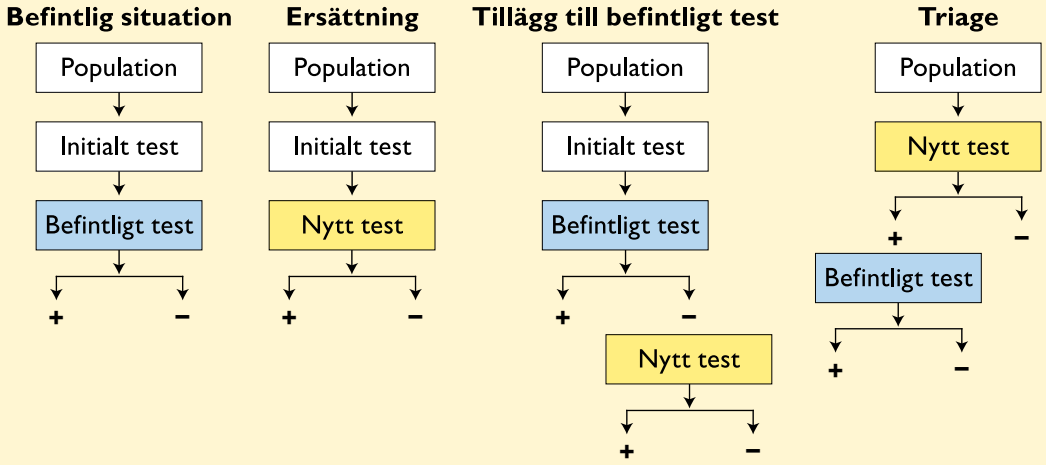
Punkterna 3–5 i modellen motsvarar Steg 4-frågan, där det kliniska värdet av ett diagnostiskt test efterfrågas, dvs ”blir utfallet (resultat till följd av testutfallet) bättre för patienter som genomgår testet jämfört med liknande patienter som inte genomgår testet?”

Andra viktiga frågor att överväga är:

- Avser översikten att undersöka den tekniska tillförlitligheten hos ett test?
- Vilka kliniska situationer (t ex primärvård, specialistvård, sjukhusvård) är relevanta?
- Är syftet att undersöka tester för screening, eller avser översikten diagnostik av patienter med misstänkt sjukdom?
- Är avsikten att undersöka tester som används för att diagnostisera etablerad sjukdom (eller tillstånd), eller vill man undersöka tillförlitligheten hos ett test (tester) för att identifiera tidiga tecken på sjukdom eller gradera sjukdomsstadier (eller samtliga)?
- Ska testet(erna) användas för att utvärdera progression av ett sjukdomsförlopp (t ex funktionsnedsättning) eller bedöma prognos?
- Det kan också vara praktiskt att försöka bestämma vilken funktion ett test har i den diagnostiska processen, se Faktaruta 7.3.

Faktaruta 7.3 Vilken funktion ska ett nytt test ha?

Om ett nytt test ska undersökas, vilken funktion ska det ha i den diagnostiska kedjan? Det kan ha olika funktioner i förhållande till existerande test(er). Syftet kan vara att ersätta eller lägga till ett befintligt test. Det kan också vara ett triage-test, dvs det nya testet ska föregå existerande test(er). Det nya testets roll och position i den diagnostiska kedjan illustreras i Figur 7.3.1.



Figur 7.3.1 Testets roll och position i den diagnostiska kedjan [14].

Faktarutan fortsätter på nästa sida

Faktaruta 7.3 Fortsättning.

Befintlig situation motsvarar utgångsläget för den population som ett nytt test ska prövas på. *Initialt test* är testresultat eller annan information (t ex anamnes och klinisk undersökning) som finns tillgänglig innan befintligt test gjorts.

Ersättning är en situation, där tillförlitligheten hos ett nytt test jämförs med den hos ett befintligt test. Jämförelsen kan också gälla hur invasiva testerna är eller att jämföra kostnader. Ett exempel är mammografiscreening där befintligt test (bildgranskning med två bröstradiologer) jämförs med nytt test (bildgranskning med en bröstradiolog + datoriserad analys (”computer-aided detection”)).

Tillägg till befintligt test innebär i exemplet att patienter med ett negativt testresultat på det befintliga testet undersöks med ett nytt test (tillägg till befintligt test). Ett exempel är positron emissionstomografi för att upptäcka metastaser, där befintligt test är datortomografi och ultraljud.

Triage är en situation där man vill undersöka tillförlitligheten hos ett test före initiala tester/befintliga tester för att kunna sortera bort individer från fortsatt testning. Ett exempel är de så kallade ”Ottawa ancle rules” (enkel klinisk undersökning av patienter med fot- och/eller ankelsmärtor), som har mycket låg andel falska negativa fynd, och som därför ofta används för att reducera antalet onödiga röntgenundersökningar för att bekräfta eller utesluta fraktur.

Förutom information om var i den diagnostiska kedjan det nya testet ska ingå, är det viktigt att veta vilka tester eller annan information som finns tillgängliga i den undersökta populationen, innan tillförlitligheten hos ett nytt test ska undersökas. Om t ex endast patienter med ett visst positivt testresultat ingår då det nya testet ska göras, har sammansättningen i populationen förändrats. Sannolikheten för sjukdom efter ett positivt testutfall ökar, vilket påverkar det nya testets sensitivitet och specificitet, se ”Rätt patientgrupp” i QUADAS-mallens fråga 1 (Bilaga 4).

Vad betyder ett diagnostiskt test för patienten?

Utvärdering av den diagnostiska tillförlitligheten hos ett test är en viktig komponent för att bestämma användbarheten hos ett test, men det kliniska värdet ligger i att förbättra patientens hälsotillstånd. Utfallet av ett diagnostiskt test kan påverka behandling, behandlingsresultat och patientens välbefinnande (ett testresultat i sig kan påverka patienten både emotionellt och beteendemässigt). Men till skillnad mot interventionsstudier är resultat av diagnostiska test intermediärt, dvs det kan påverka men bestämmer inte direkt behandlingsutfallet hos patienten. Det kliniska värdet, dvs hur resultatet av ett test påverkar det kliniska beslutsfattandet, effekten på patientens välbefinnande och/eller på utfallet av en påföljande behandling är därför viktiga steg vid utvärdering av diagnostiska test. Ett test med god tillförlitlighet kan, men behöver inte vara, effektivt

och till nytta för patienten. Studier som undersöker värdet av en diagnostisk intervention är dock sällan tillgängliga, speciellt inte för nya tester. Dessutom saknas tillräckligt bra och/eller oberoende referensstandard för flera sjukdomstillstånd. Då kan ett tests värde bara bedömas genom att studera naturalförloppet eller utfallet av en behandling.

Idealiskt bör SBU studera såväl den diagnostiska metoden som efterföljande behandlingsalternativ för att kunna bedöma patientnytta. Eventuellt kan man söka efter systematiska översikter för de behandlingsmetoder som är aktuella och ta upp dem i den avslutande diskussionen om potentiell nytta. Observera att även om översikten bara behandlar en begränsad del av den diagnostiska processen, måste aspekter på nyttan och värdet av ett test kommenteras och diskuteras i rapporten.

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, referenstest (motsvarande ”control”), ”outcome”). Exempel på inklusionskriterier uppställda enligt PICO från SBU-rapporten ”Rotfyllning” visas i Exempel 7.2.

Det lönar sig att noga tänka igenom inklusions- och exklusionskriterier innan man går vidare med granskningsdelen. Det gäller t ex 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, och vilken/vilka referenstester som ska accepteras.

Exempel 7.2 Inklusionskriterier formulerade som PICO [15].

PICO	Inklusionskriterier
Population	Patienter som kan förväntas få undersökningen eller testet i klinisk praxis. Permanenta tänder
Indextest	Kliniska symtom, annan klinisk information, kliniska tester eller biologiska markörer testade mot en referensstandard
Referenstest (”control”)	För pulpastatus i vital vävnad: histologisk undersökning alternativt symtom och klinisk/röntgenologisk information vid prospektiv studiedesign För vitalitetsbestämning av pulpa: som ovan eller inspektion/sondering av pulparummet, alternativt röntgenologisk undersökning med fortsatt rotutveckling hos tänder med oavslutad rotutveckling
Utfall (”outcome”)	Frisk pulpa/pulpainflammation eller pulpadöd

Granskning av studier som uppfyller inklusionskriterierna

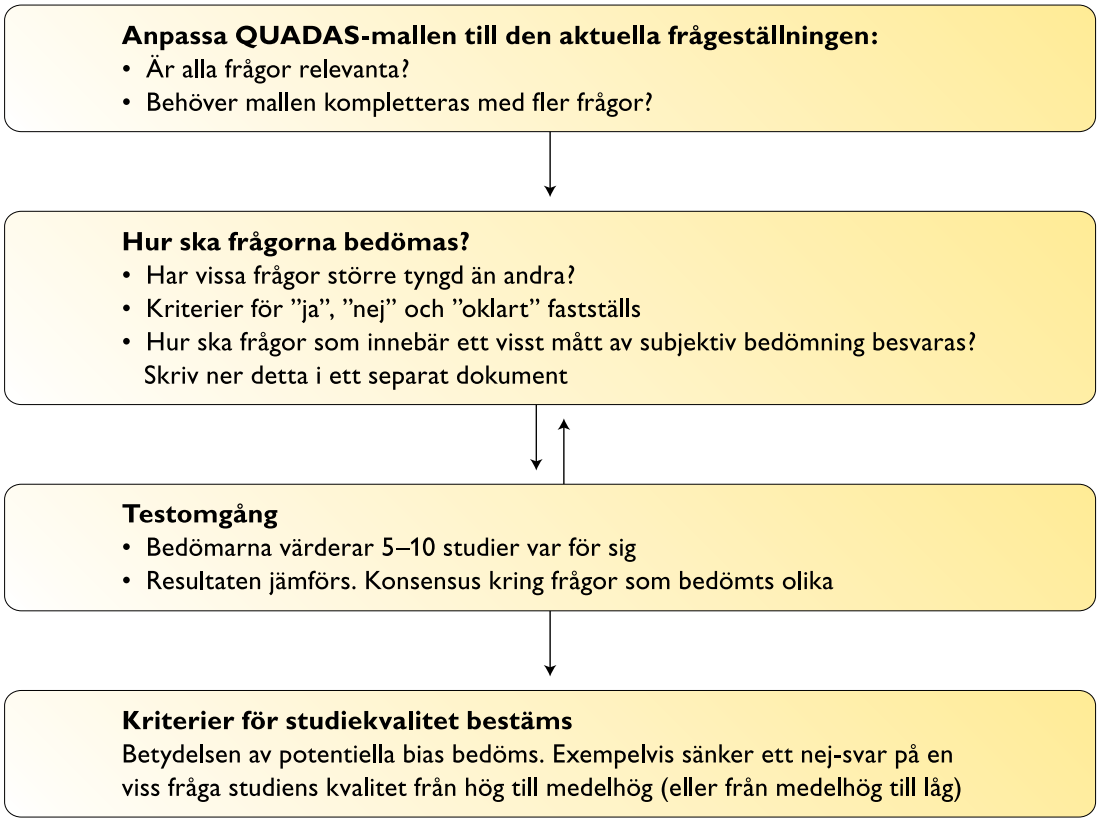
STARD (STAndards for the Reporting of Diagnostic accuracy studies) [16] innehåller rekommendationer för hur studier om diagnostisk tillförlitlighet bör designas, genomföras och rapporteras. STARD består av 25 delmoment med kvalitetsaspekter som i flera avseenden är desamma som för interventionsstudier.

Granskningsmall (Bilaga 4)

SBU använder en granskningsmall enligt QUADAS [2]. Det är den mall som används mest internationellt och som korresponderar väl med STARD. Den underliggande principen för QUADAS är att bristfälliga metodologiska karakteristika kan introducera bias eller begränsa en studies applicerbarhet. Mallen fungerar som en checklista, och den reviderade versionen [17] består av 11 frågor om en studies metodologiska kvalitet och applicerbarhet, se Bilaga 4. Varje fråga har tre svarsalternativ: ”ja”, ”nej” eller ”oklart”. Hur dessa svarsalternativ ska tolkas och användas framgår också av QUADAS [2,17]. QUADAS-mallen är i första hand anpassad för att bedöma tvärsnittsdata. Karaktär och typ av frågeställningar varierar dock mellan olika översikter. Det kan innebära att någon/några frågor inte är relevanta eller behöver omformuleras, och/eller att mallen behöver kompletteras med fler frågor. Mallen tar t ex inte speciell hänsyn till studier som jämför flera tester, och den behöver kompletteras med ytterligare kvalitetsfrågor när det krävs longitudinell uppföljning av patienten för att verifiera en diagnos. Exempel på andra frågor som kan behöva läggas till är:

- Var studiens syfte(n) definierad(e) innan studiens start?
- Var tröskelvärden (”cut-off”) bestämda innan studien startade?
- Är teknologin för indextestet oförändrad sedan studien gjordes?
- Ger studien en klar definition på vad som ansågs vara ett ”positivt” resultat?
- Hade de som utförde test(erna) adekvat utbildning/träning?
- Var behandling uppskjuten tills både indextest och referenstest hade utförts?
- Rapporterades observatörsvariation och var den inom acceptabla gränser?
- Var kommersiella intressen involverade i finansieringen av studien?

Flera av frågorna i QUADAS kräver ett visst mått av subjektiv värdering, medan andra är mer ”svart-vita”. För de mer subjektiva frågorna, som bedömning av patientsammansättning, är det viktigt att formulera klara riktlinjer för hur frågan ska bedömas med utgångspunkt från den aktuella frågeställningen. Det är lämpligt att bedömarna, oberoende av varandra, gör en pilot-checklista på ett urval av minst fem studier och därefter diskuterar eventuella diskrepanser. Graden av överensstämmelse och viktiga diskrepanser bör dokumenteras. Ett sätt att säkerställa att granskarnas bedömning av studierna har tillräckligt god överensstämmelse är att beräkna kappa. De slutgiltiga bedömningsgrunderna sammanställs i en manual. Flödesschema för granskning med QUADAS ges i Figur 7.1.



Figur 7.1 Flödesschema för granskning av diagnostiska studier.

Bedömning av studiekvalitet

Den metodologiska kvaliteten hos en studie bedöms utifrån risker för bias (risk för systematiska fel beroende på problem med studiedesign, genomförande och/eller rapportering). Här värderar vi studiens interna validitet. Den externa validiteten bestäms ofta av studiens val av patientspektrum. Är resultaten applicerbara på den population(er) som översiktens frågeställning avser? En studie kan ha god metodologisk kvalitet, men relevansen kan ifrågasättas vid närmare bedömning av studien. En studies relevans bedöms alltså i separat ordning.

Riskerna för bias är delvis desamma för diagnostiska studier och interventionsstudier. Det gäller t ex 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). Men i motsats till behandlingsstudier, där randomisering är vanlig, utgör populationen i diagnostiska studier ofta konsekutivt insamlade individer som genomgår såväl indextest som referenstest. De viktigaste källorna till bias i diagnostiska studier framgår av Tabell 7.2. Som framgår av tabellen, finns en rad förhållanden, både i studiedesign och genomförande, som kan leda till bias eller variation. Det är dock osäkert hur stor den verkliga effekten av dessa bias är [18].

Tabell 7.2 Källor till bias i studier om diagnostisk tillförlitlighet [3,18].

Typ av bias	När inträffar det?	Under- eller överskattning av diagnostisk tillförlitlighet
Patienter		
Patientsamman-sättningsbias	När inkluderade patienter inte representerar det avsedda spektrum av allvarlighetsgrad hos det sökta tillståndet	Beror på skillnaden mellan det sökta tillståndet och det spektrum som inkluderats i studien
Selektionsbias	När avsedda patienter 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 tillgång till information om kliniska data som ålder, kön och symtom finns tillgänglig då index-testet tolkas (gäller framför allt röntgenbilder)	Leder till högre sensitivitet men har liten påverkan på specificitet
Referensstandard		
Felklassifikationsbias	När referensstandard inte korrekt klassificerar patienter med det sökta tillståndet	Beror på om båda testerna gör samma misstag
Partiell verifikationsbias	När ett icke randomiserat urval av patienter inte genomgår referensstandardtestet	Leder ofta till överskattning av sensitivitet. Effekt på specificitet varierar
Differentiell verifikationsbias	När en del av patienterna verifieras med ett andra eller tredje referens-standard, speciellt om denna selektion beror på resultat av indextestet	Leder ofta till överskattning
Inkorporationsbias	När indextestet inkorporeras i en (sammansatt) referensstandard	Leder ofta till överskattning
Sjukdoms-progressionsbias	När patientens tillstånd förändras mellan administrering av indextest och referensstandard	Över- eller underskattning beroende på förändringen i patientens tillstånd
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

Bedömningen görs med utgångspunkt från i vilken grad olika eventuella potentiella bias bedöms påverka studiens validitet. QUADAS-mallen ska ses som ett verktyg och hjälpmedel vid bedömning av studiekvalitet. Samstämmig bedömning av varje fråga bör eftersträvas. Om bedömarna inte är överens bör en ytterligare utslagsgivande bedömare tillfrågas. Därefter bestäms vilka krav som ska ställas för att en studie ska klassas som hög, medelhög eller låg kvalitet.

Ett exempel på hur QUADAS kan behöva anpassas är från projektet om diagnostik vid affektiv sjukdom, som undersöker tillförlitligheten hos olika skattningsskalor för depression. Här beslöts att lägga till en kolumn för ”specifika krav”. För hög studiekvalitet krävdes t ex att patienterna inte fick vara annonsrekryterade (Fråga 1).

Man bör eftersträva att följa granskningsmallen, och om man lägger till frågor, bör dessa vara så allmängiltiga som möjligt. Det bästa sättet att åstadkomma det, är att så långt möjligt specificera inklusions- och exklusionskriterier. Det är t ex praktiskt att bestämma vilka populationer som ska inkluderas. Ska fall–kontrollstudier accepteras eller enbart studier som använder konsekutivt valda patienter med misstänkt sjukdom? Ett annat exempel är att specificera vilken eller vilka referensstandard som ska accepteras. Brister i rapportering är vanligt i diagnostiska studier. Avsaknad av för den aktuella frågeställningen viktig information kan vara ett exklusionskriterium. Hur ”välvillig” man ska vara avseende inklusions- respektive exklusionskriterier beror ofta på det aktuella kunskapsläget för den aktuella frågeställningen.

Dataextraktion och tabellering

Data från de studier som bedömts ha hög eller medelhög kvalitet tabelleras på samma sätt som för interventionsstudier. Tabell 7.3 visar tabellhuvud med rubriker för diagnostiska test. För den aktuella frågeställningen relevanta parametrar kan läggas till under en passande kolumnrubrik.

Tabell 7.3 Tabellhuvud för studier om diagnostisk tillförlitlighet.

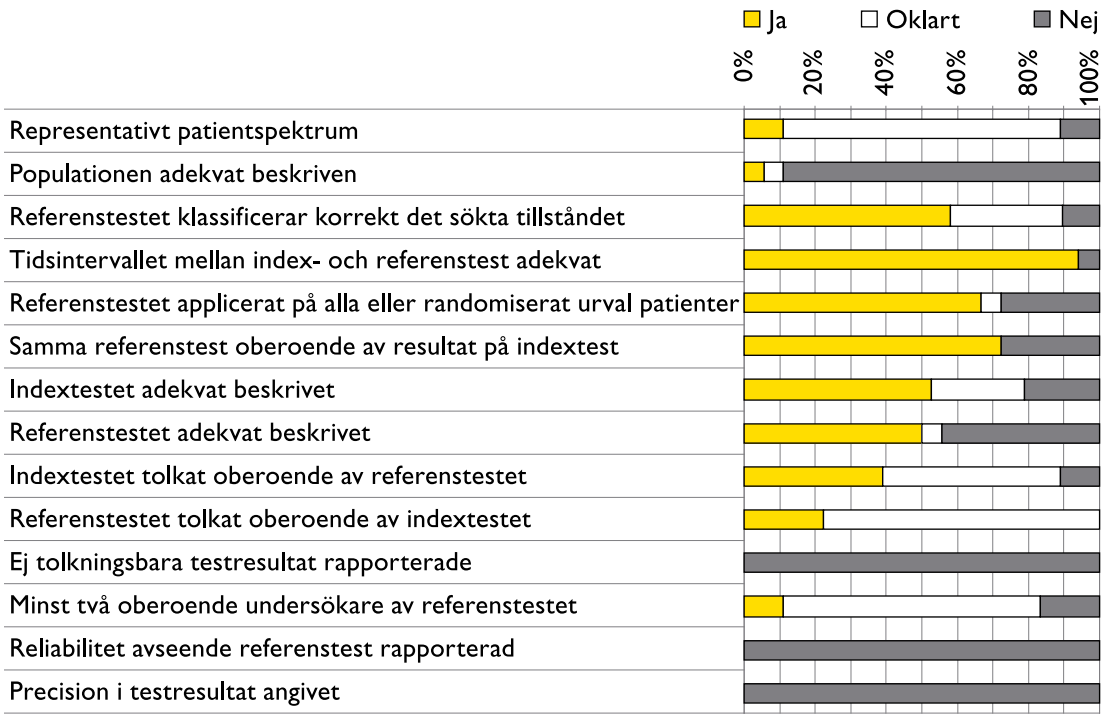
First author Year Country Reference	**Aim**	**Study design Population characteristics Setting**	**Index test**	**Reference test**	**Results**	**Study quality Relevance Comments**

Under ”Aim” redovisas författarnas syfte(n) med studien. Åtminstone initialt i arbetet med en tabell är det värdefullt att tydliggöra dessa. Det kan t ex klargöra om de data man extraherar från en artikel motsvarar författarnas ursprungliga syften, eller om

data endast är rapporterade bifynd. Under ”Study design” redovisas vilken design som använts (t ex tvärsnittsstudie, kohortstudie). Här beskrivs också den undersökta patientpopulationens karakteristika, och hur den rekryterades. ”Setting” avser under vilka omständigheter patienterna rekryterades (t ex primärvård, specialistvård, sjukhusvård). Under ”Index test” beskrivs kortfattat vad indextestet(n) bestod av. Detsamma gäller ”Reference test”. Under ”Results” ska kvantitativa mått presenteras. Utöver det kan en kort textversion utifrån författarnas slutsatser också vara lämplig. Under ”Study quality/ Relevance/Comments” anges studiekvalitet och huvudsaklig förklaring till varför en studie nedgraderats i kvalitet/relevans.

Konfidensintervall för sensitivitet och specificitet bör anges. Om dessa inte rapporteras i originalstudien, men går att beräkna görs detta. En asterisk (*) framför beräkningen anger att dessa data inte redovisats i originalstudien.

Ett vanligt problem vid bedömning av diagnostiska studier är bristande kvalitet i både studiedesign och rapportering [18,19]. Med utgångspunkt från QUADAS-kriterierna bör en sammanställning av studiernas kvalitet ingå i rapporten. Det underlättar det fortsatta arbetet med att bedöma GRADE (Kapitel 10). Ett exempel på en sådan sammanställning visas i Figur 7.2. Det ger en snabb översikt över var det kan finnas problem kring studiekvalitet. Vilket problem som är viktigast att värdera kan variera från fråga till fråga.



Figur 7.2 Rapportering av 14 kvalitetskriterier (modifierade efter QUADAS-kriterier) hos 18 studier avseende diagnostik av tandpulpa. Procentuell fördelning av ”ja”, ”oklart” och ”nej” för respektive kriterium [15].

Därför rekommenderar SBU inte någon standardiserad poängsättning. I det aktuella projektet kan man med fördel dokumentera vilka faktorer som vägt tyngst vid värdering av studiekvalitet.

Statistisk analys och sammanställning av data

Analys av data från studier om diagnostisk tillförlitlighet skiljer sig på flera sätt från analys av studier som avser terapeutisk intervention:

- Diagnostisk tillförlitlighet kvantifieras vanligtvis med *två* mått: sensitivitet och specificitet, som inte kan reduceras till ett enkelt summationsmått (som diagnostiskt ”odds ratio”) utan att man förlorar information. Två andra mått är positivt och negativt prediktionsvärde.

- För att bestämma när ett test är positivt krävs ofta att man väljer ett tröskelvärde, t ex för biokemiska test.

- Studier är oftast heterogena, vilket gör att sammanställning av data kan vara problematisk.

Generellt sett är metoderna för statistisk syntes av diagnostiska studier inte lika väl utarbetade som för interventionsstudier. Man bör starta med enkel deskriptiv sammanställning av data. Det kan göras med *kopplade ”forest plots”* och en *enkel summations-ROC-kurva*. Om man har tillräckligt många rimligt homogena studier kan bivariata/hierarkiska modeller för metaanalys användas.

Analys av heterogenitet

Det finns flera källor till heterogenitet hos diagnostiska studier [20]. Den kan bero på slumpen, men oftast handlar det om äkta heterogenitet, där en vanlig orsak är att olika tröskelvärden (”cut-offs”) använts för att definiera ett positivt (eller negativt) testresultat. Andra orsaker kan vara skillnader i patientsammansättning (allvarlighet hos sjukdomen eller samsjuklighet) eller så kallad partiell verifikationsbias (ett icke randomiserat urval av patienter genomgår inte referenstestet). Olika teknologi för indextest och/eller referenstest, skillnader mellan observatörer, olika studiedesign och genomförande kan också orsaka heterogenitet hos studier. Det finns nästan alltid heterogenitet hos diagnostiska studier, och *orsakerna* bör alltid analyseras. Om heterogeniteten är stor, kommer en statistisk syntes inte att vara meningsfull.

Ett första steg bör vara att bedöma *graden* av heterogenitet. Den kan visualiseras grafiskt med hjälp av ”forest plots”, där man har två kopplade ”plots”, dvs en för sensitivitet och en för specificitet (Exempel 7.3). ”Forest plots” kan också göras för positivt och negativt prediktionsvärde och för ”likelihood”-kvoter. Exempel på kopplade ”forest plots” och enkel sammanställning av studieresultat ges i Exempel 7.3.

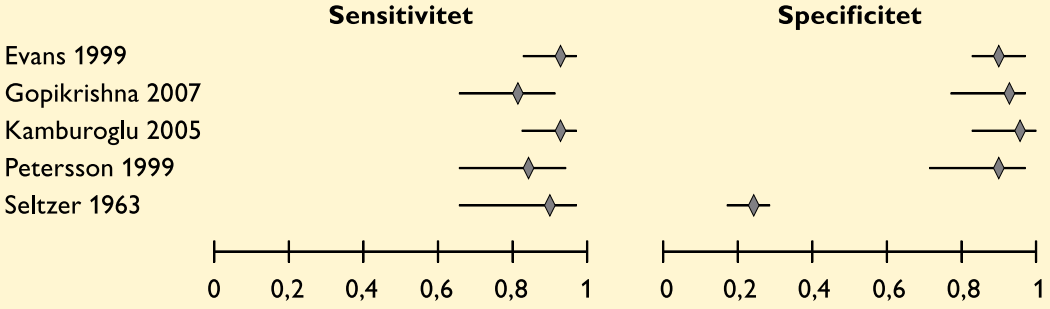
Forest plottarna liksom sammanställningen i Exempel 7.3 visar att heterogeniteten för specificitet är hög, och det är uppenbart att en studie avviker från de övriga med låg specificitet (Seltzer). Här finns flera orsaker till heterogenitet. Rapportering av hur patienterna rekryterats var bristfällig i flera av studierna, och referensstandarden i en studie var en annan än i de övriga (Seltzer). *Här är det alltså olämpligt att göra en metaanalys (där värden för sensitivitet respektive specificitet vägs samman till ett mått).* En möjlighet kan vara att göra metaanalys på en subgrupp med rimligt homogena studier.

Deskriptiv sammanställning av heterogena studier

Förutom den sammanställning som illustreras i Exempel 7.3, kan resultat från heterogena studier sammanställas grafiskt med en ROC-kurva (”receiver operator characteristic”). Den visar hur resultaten fördelar sig, men ger inte precisionen i enskilda studier. Värdering av heterogenitet går heller inte att göra. Ett exempel visas i Figur 7.3.

Kopplade ”forest plots” och enkla ROC-kurvor kan göras i Cochranes Review Manager, RevMan (http://ims.cochrane.org/revman). Om den inledande analysen visar att studierna är alltför heterogena, bör man inte gå vidare med mer avancerade statistiska metoder.

Exempel 7.3 Kopplade ”forest plots” av sensitivitet och specificitet samt sammanställning av sensivitet och specificitet av fem studier som undersöker tillförlitligheten hos kyla-test (en pellet doppad i etylklorid appliceras på tandytan) för att bestämma om en tandpulpa är vital eller non-vital.



Exemplet fortsätter på nästa sida

Exempel 7.3 Fortsättning.

Sammanställning av studierna

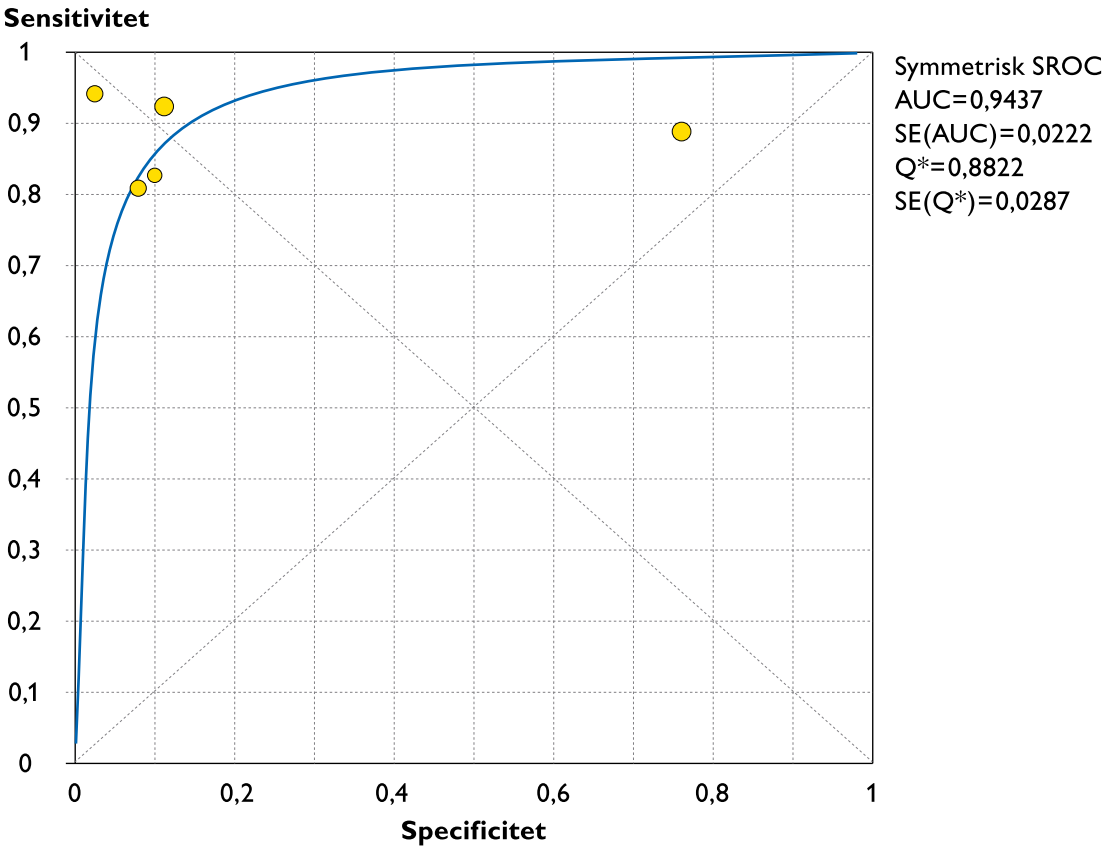
Sensitivitet				
Studie	*Sensitivitet*	*(95% KI)*	*SP/(SP+FN)*	*SN/(SN+FP)*
Evans 1999	0,92	0,82; 0,98	49/53	72/81
Gopikrishna 2007	0,81	0,66; 0,91	34/42	35/38
Kamburoglu 2005	0,94	0,84; 0,99	49/52	40/41
Petersson 1999	0,83	0,64; 0,94	24/29	27/30
Seltzer 1963	0,89	0,65; 0,99	16/18	29/121
Specificitet				
Studie	*Specificitet*	*(95% KI)*	*SP/(SP+FN)*	*SN/(SN+FP)*
Evans 1999	0,89	0,80; 0,95	49/53	72/81
Gopikrishna 2007	0,92	0,79; 0,98	34/42	35/38
Kamburoglu 2005	0,98	0,87; 0,99	49/52	40/41
Petersson 1999	0,90	0,73; 0,98	24/29	27/30
Seltzer 1963	0,24	0,17; 0,33	16/18	29/121

FN = Falska negativa; FP = Falska positiva; KI = Konfidensintervall; SN = Sanna negativa; SP = Sanna positiva

Statistisk syntes av rimligt homogena studier

Om studierna är rimligt homogena [21], dvs patientpopulation, indextest och referenstest är jämförbara, och det inte finns uppenbara risker för allvarliga bias i studiedesign och genomförande i de enskilda studierna, kan sensitivitet, specificitet och ”likelihood”-kvoter kombineras. Enkla metoder som kopplade ”forest plots” och ROC tar inte hänsyn till den negativa korrelationen mellan sensitivitet och specificitet. Resultaten kan istället vägas samman med hjälp av någon av nedanstående modeller.

Hierarkiska modeller för statistisk analys kan användas om man har tillräckligt många rimligt homogena studier. *Bivariat metaanalys* tar hänsyn till den underliggande korrelationen mellan sensitivitet och specificitet [22]. *Hierarkisk summa ROC-analys* baseras på logaritmen för diagnostisk oddskvot (DOR) och tar hänsyn till bl a tröskelvärde. Dessa båda metoder anses vara mer robusta och bättre lämpade för metaanalys av diagnostiska data [3,23], men de kräver goda kunskaper om modellerna och tillgång till statistisk programvara av typen Stata, SAS eller SPSS (senare versioner). Dessa statistiska modeller behandlas utförligt i Cochranes handbok, Kapitel 10 [24].



Figur 7.3 ROC-kurva som summerar de fem studierna som undersökte tillförlitligheten hos kyla-test för att bestämma om en tandpulpa är vital eller non-vital. En studie avviker från de övriga genom att ha hög andel falska positiva resultat (Seltzer).

Resultattabell

Förutom sedvanlig tabellering av inkluderade studier ska viktiga resultat sammanställas i en tabell enligt GRADE (Kapitel 10). Ett exempel från en SBU Alert-rapport om mammografi [25] visas i Tabell 7.4.

Patientnytta

Även om det går att sammanställa resultat i t ex en GRADE-tabell, saknas oftast utfallsmåttet ”patientnytta”. Det innebär att man i diskussionen behöver påtala denna svaghet. Se vidare i Kapitel 10.

Kostnadseffektivitet

Kostnadseffektiviteten för diagnostiska studier redovisas i Kapitel 11.

Tabell 7.4 Tillförlitligheten vid screening med mammografi för cancerdiagnostik jämfördes hos två metoder: dubbelgranskning med två radiologer versus enkelgranskning (en radiolog) + CAD (”computer-aided detection”) [25]. Referensstandard var biopsi av bröstvävnad eller uppföljning av patienten. Tabellen är samma typ som den som används för GRADE.

Effekt-mått	**Antal patienter (antal studier)**	**Sant positiva: Enkelgransk-ning + CAD (95% KI)**	**Sant positiva: Dubbel-granskning (95% KI)**	**Absolut skillnad (95% KI)**	**Veten-skapligt under-lag**	**Kommen-tarer***
Cancer-detek-tions-frekvens	28 204 (1)	0,702% (0,6; 0,8)	0,706% (0,6; 0,8)	0,004% (Ej statistiskt signifikant skillnad)	⊕○○○ Otill-räckligt	Brister i studie-kvalitet –1 Begränsad överförbar-het –1 Oprecisa data –1
Effekt-mått	**Antal patienter (antal studier)**	**Andel återkallade: Enkelgransk-ning + CAD (95% KI)**	**Andel åter-kallade: Dubbel-granskning (95% KI)**	**Absolut skillnad (95% KI)**	**Veten-skapligt under-lag**	**Kommen-tarer***
Åter-kallnings-frekvens	28 204 (1)	3,9% (3,7; 4,1)	3,4% (3,2; 3,6)	0,5% (0,3; 0,8)	⊕○○○ Otill-räckligt	Brister i studie-kvalitet –1 Begränsad överförbar-het, endast en studie –2

CAD = ”Computer-aided detection”; KI = Konfidensintervall

** Brister i studiekvalitet = Risk för bias, dvs sensitivitet troligtvis övervärderad (ofullständig uppföljning av kvinnor med negativt testresultat).*
Begränsad överförbarhet = Endast bröstradiologer med lång klinisk erfarenhet ingick i studien.
Oprecisa data = Vida konfidensintervall för skillnaden i sensitivitet mellan dubbelgranskning och enkelgranskning + CAD.

Referenser

1. Wilson JMG, Jungner G. Principles and practice of screening for disease. Geneva: WHO; 1968. http://whqlibdoc.who.int/php/WHO_PHP_34.pdf

2. 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.

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

4. Westwood ME, Whiting PF, Kleijnen J. How does study quality affect the results of a diagnostic meta-analysis? BMC Med Res Methodol 2005;5:20.

5. Fletcher RH, Fletcher SW. Clinical epidemiology. The essentials. Philadelphia: Lippincott Williams & Wilkins; 2005.

6. 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.

7. 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.

8. 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.

9. Paulus WJ, Tschöpe 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.

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

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

12. 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.

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

14. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006;332:1089-92.

15. SBU. Rotfyllning. En systematisk litteraturöversikt. Stockholm: Statens beredning för medicinsk utvärdering (SBU); 2010. SBU-rapport nr 203. ISBN 978-91-85413-39-3.

16. 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.

17. Reitsma JB, Rutjes AWS, Whiting P, Vlassov VV, Leeflang MMG, Deeks JJ. Chapter 9: Assessing methodological quality. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.0. The Cochrane Collaboration, 2009. http://srdta.cochrane.org/

18. 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.

19. Lijmer JG, Mol BW, Heisterkamp S, Bonsel 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.

20. Dinnes J, Deeks J, Kirby J, Roderick P. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. Health Technol Assess 2005;9:1-113.

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

22. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005;58:982-90.

23. Harbord RM, Whiting P, Sterne JA, Egger M, Deeks JJ, Shang A, et al. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. J Clin Epidemiol 2008;61:1095-103.

24. Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Chapter 10: Analysing and Presenting Results. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 0.9.0. The Cochrane Collaboration, 2010. http://srdta.cochrane.org/

25. SBU. Datorassisterad granskning inom mammografiscreening (CAD). Stockholm: Statens beredning för medicinsk utvärdering (SBU); 2011. SBU Alert-rapport nr 2011-05. ISSN 1652-7151. http://www.sbu.se

8. **Värdering och syntes av studier utförda med kvalitativ analysmetodik**

VERSION 2012:2

Det är vanligt att undersökningar av t ex en persons livskvalitet görs med kvantitativa metoder i form av olika skattningsformulär. Om man vill skapa en djupare förståelse för en persons subjektiva upplevelse av t ex livskvalitet, personers individuella upplevelser, erfarenheter, uppfattningar och handlingar kan det vara mer relevant att undersöka detta med kvalitativa forskningsmetoder. Detta för att förstå sammanhanget från personens synvinkel. En viss överlappning av beskrivningarna kan ske eftersom frågorna som används i skattningsformulären vanligen är baserade på kvalitativ metodik i form av analys av t ex intervjuer, men i huvudsak tillkommer helt ny information. I detta fall kompletterar alltså metoderna varandra.

Ett exempel är en dansk HTA-rapport om uppföljning av patienter med cancer i vilken studier utförda med såväl kvantitativ som kvalitativ metodik utvärderades. Studierna utförda med kvalitativ metodik visade att patienter som behandlats för cancer tycker att kontroller bidrar till livskvalitet genom att ge trygghet, bekräftelse och lättnad om undersökningen var utan anmärkning. De kvantitativa studierna med skattningsskalor visade ingen skillnad mellan patienter som fått uppföljning jämfört med patienter som inte fått det. De kvalitativa studierna fångade alltså upp information ”mellan” frågorna i skattningsformulären [1].

SBU utvärderar metoder inom hälso- och sjukvården och som en del av detta granskas även hur patienten eller patientens anhöriga upplever olika aspekter inom vården, som erfarenheter av behandling, diagnostik eller av att leva med olika hälsotillstånd. Detta kapitel fokuserar därför på kvalitativ forskning i förhållande till patientupplevelser.

Kvalitativ forskning innebär någon form av forskning vars resultat inte är produkter av statistiska processer eller andra kvantitativa ansatser. Den erbjuder insikt i sociala, emotionella och experimentella fenomen. Syftet är att få en uppfattning, att nå en förståelse, att utforska särdrag i olika miljöer och kulturer samt att förstå relationen mellan olika processer. Kännetecknande för kvalitativ forskningsmetodik är också att forskaren ofta själv är ett redskap för datainsamling. De flesta kvalitativa studier fokuserar på en företeelse, eller ett litet antal företeelser. Studierna tillför en detaljrik och djup kunskap, som kan ge en ökad förståelse för fenomen ur ett perspektiv som kvantitativa fynd inte kan fånga. För mer information om skillnaderna mellan kvantitativ och kvalitativ forskning, se Tabell 8.1.

Tabell 8.1 Jämförelse av kvantitativ och kvalitativ forskning [2].

	Kvantitativ forskning	**Kvalitativ forskning**
Vad är verklighet?	• Verkligheten existerar oberoende av människors tro och tolkningar och kan mätas direkt (positivism)	• Det finns en oberoende verklighet men den kan bara bli tillgänglig via mänsklig tolkning, vilket leder till flera perspektiv (interpretivism) • En del kvalitativa forskare hävdar att det inte finns någon oberoende verklighet, bara individuella eller delade sociala konstruktioner
Förhållandet mellan forskaren och deltagaren	• Forskaren påverkar ej data från deltagaren. Objektiv forskning utan värderingars inverkan anses vara möjlig	• Även om forskaren ämnar vara så neutral som möjligt, är det ofrånkomligt att forskaren och deltagaren påverkar varandra. Dataanalysen formas av forskarens värderingar och en objektiv forskning fri från värderingar är inte möjlig
Kunskapens förvärvande	• Mestadels genom deduktion	• Induktion och deduktion sker under flera steg av forskningsprocessen
Frågeställningar	• Hur mycket? • Hur många? • Finns det en statistisk skillnad mellan…? • Finns det en korrelation? • Vilka är de starkaste prediktorerna för...?	• Vad? • Hur? • Varför?
Ansats	• Reduktionistisk • Studiedesignen bestäms i förväg • Förståelse för ett fenomen från forskarens perspektiv • Fokus på objektiv mätning	• Holistisk • Studiedesignen kan vara flexibel för att möjliggöra fortsatta utforskningar av idéer som kommer upp • Förståelse för ett fenomen från deltagarens perspektiv • Fokus på subjektiv mening, förståelse och process
Sammanhang (kontext)	• Vikten av sammanhanget varierar men kontextuella faktorer är ofta eliminerade i kontrollerade experimentella studier	• Sammanhanget är viktigt i formandet av mening och förklaringar • Forskningen äger rum i naturliga miljöer
Forskningsinstrument	• Validerat instrument, mått, skattningsskala eller enkät	• Forskaren är det primära forskningsinstrumentet
Urval	• Randomiserat urval eller sannolikhetsurval • Representativ population • Förbestämt på basen av poweruträkning	• Strategiskt eller teoretiskt urval • Avspeglar populationens mångfald • Tillräckligt flexibel för att kunna drivas av framkommen teori • Urvalets storlek bestäms i idealfallet av datamättnad

Tabellen fortsätter på nästa sida

Tabell 8.1 Fortsättning.

	Kvantitativ forskning	Kvalitativ forskning
Data	• Siffror	• Ord och bilder
Analys	• Statistisk analys • Analysenheterna är variabler • Analysen sker efter datainsamlingen	• Icke-statistisk analys • Analysenheterna är teman • Analysen och datainsamlingen kan ske samtidigt
Utfall	• Deskriptiv statistik • Statistisk evidens för korrelation/skillnader mellan grupper • Förutsägelse av den oberoende variabelns effekt på den beroende variabeln	• Detaljerad beskrivning • Klassificering • Typologier • Förståelse
Generaliserbarhet/överförbarhet	• Probabilistisk • Inferentiell	• Representativ • Inferentiell • Teoretisk

Kvalitativ forskning kan användas för att undersöka personers uppfattningar, erfarenheter, upplevelser och mening i relation till ett visst fenomen. Den är även värdefull då man försöker förstå vilka faktorer som inverkar på systemförändringar samt på personers benägenhet att förändra sig. Den gemensamma nämnaren för kvalitativ forskning är att forskaren söker en förståelse och vill skapa en allmängiltig bild av det som undersöks, och ibland generera teori.

En del av forskningsfältet kring hälso- och sjukvård som har utvecklats under de senaste decennierna är forskning om hur hälso- och sjukvården organiseras. Organisationsforskningen har blivit allt viktigare eftersom olika aktörer har insett att frågor kring hur hälso- och sjukvården organiseras, styrs och implementeras kan påverka hur framgångsrikt en metod införs och används i vården.

Forskningsresultatens användbarhet

När man gör en systematisk översikt måste man i förväg bestämma hur forskningsresultaten ska användas. Beslutet påverkar arbetsprocessen, speciellt med tanke på syntes. Nedan följer exempel på hur man kan använda sig av kvalitativa forskningsresultat.

Ett icke-systematiskt tillvägagångssätt för att använda sig av kvalitativa forskningsresultat är att i diskussionen använda fynden från en eller flera kvalitativa studier för att tolka och stödja resultaten från kvantitativa studier. Detta för att öka förståelsen, eller för att placera de kvantitativa resultaten i ett sammanhang.

Ett annat alternativ är en formell syntes av de kvalitativa fynden. Några år efter att kvantitativ metaanalys etablerats som en metod inom samhällsvetenskaplig forskning presenterades en motsvarande metod för kvalitativ syntes. Upphovsmännen, Noblit och Hare, benämnde denna syntesmetod för metaetnografi. Under åren som gått har olika syntesmetoder med varierande benämningar presenterats, men grundtanken är ofta densamma, dvs att en översikt görs av de kvalitativa fynden, enskilt eller vid sidan av den kvantitativa syntesen [3]. Den kvalitativa syntesen används för att tolka resultaten av den kvantitativa syntesen, eller för att besvara frågeställningar som den kvantitativa analysen inte ger svar på [4]. Denna syntes liknar GRADE-systemet som används för kvantitativa studier i och med att data från olika studier syntetiseras per effektmått (Kapitel 10). För exempel på kvalitativ syntes, se avsnittet ”Syntes” i detta kapitel.

Bland de internationella HTA-organisationerna har intresset för att utvärdera kvalitativ forskning ökat efter att man insett att HTA inte alltid enbart handlar om effekt. HTA rör även aspekter kring varför och hur metoder fungerar, etiska dilemman, hur patienter och allmänheten relaterar till en given metod, samt vilka krav en metod ställer i termer av kunskap och färdighet för fackmän och organisationer. Genom att göra en syntes av kvalitativa studier i samband med HTA kan man erbjuda beslutsfattare det bästa möjliga evidensbaserade underlaget för att exempelvis kunna bedöma patientaspekter kring införandet av en metod. Detta underlag kan också utgöra ett stöd för olika prioriteringsgrupper på både lokal, regional och nationell nivå. Även påbörjandet av nya, kostsamma och obehövliga studier kan undvikas, dvs man undviker att göra fler primärstudier där evidens redan finns. På så sätt kan man även undvika att störa exempelvis väldigt sjuka patienter med intervjuer, observationer eller enkäter [5].

Forskningsresultatens överförbarhet

Den kvalitativa forskningsmetoden har kritiserats för att vara alltför beroende av sammanhanget, för att inte ha tillräckligt många deltagare samt för att inte producera överförbara eller generaliserbara resultat. Men generaliserbarhet eller överförbarhet kan ha två dimensioner. Å ena sidan behövs kvantitativa studier baserade på ett stort urval för att kunna dra slutsatser kring hur mycket, hur ofta, hur många etc. Å andra sidan behövs mer djupgående studier som baseras på ingående undersökningar och analyser för att identifiera ett fenomen. Inom kvalitativ metod används även begreppet transferabilitet. Transferabilitet avser resultatens överförbarhet och har också definierats som likhet mellan olika sammanhang. Det finns flera olika utgångspunkter för hur man kan se på överförbarhet av kvalitativa forskningsresultat.

I en artikel av Larsson [6] presenteras följande resonemang kring överförbarhet av studieresultat. I ett första resonemang hävdas att det inte finns något behov av att överföra resultaten från en ideografisk studie (om begrepp) eftersom de är baserade på logiken att de utgör en unik del, som tillsammans med andra bidrar till en helhet. Om fokus ligger på ”negativa” fall som urholkar den etablerade sanningen, är det inte meningsfullt att överföra studieresultaten, eftersom universalitet då inte råder. Dessa två resonemang medför dock inte att dessa studier inte skulle vara meningsfulla i sin egen rätt.

Man kan dock öka potentialen för överförbarhet genom att i studien inkludera så varierande fall av samma fenomen som möjligt. Detta resonemang innebär att man inte kan överföra resultaten från ett specifikt fall eller kategori, utan från en mängd av fall. Variationen i studien förväntas då också existera i andra relevanta situationer som man vill överföra resultaten till. Detta resonemang för överförbarhet är inte tillämpbart på studier med för få deltagare, exempelvis små intervjustudier.

Ett annat resonemang fokuserar på sammanhang och på likheten mellan dessa. Fokus måste då ligga på vad som är empiriskt känt istället för teoretiska antaganden. Eftersom man måste bedöma likheten mellan sammanhanget empiriskt i efterhand, ska man bedöma när det faktiskt föreligger en likhet mellan olika sammanhang. Detta förutsätter också att det är sammanhanget som bestämmer ett fenomen eller mönster.

Det sista resonemanget bygger på att kvalitativ forskning ofta producerar tolkningar, teoretiska begrepp eller beskrivningar, alltså mönster eller konfigurationer som kan kännas igen i den empiriska världen. Detta kan ses som en slags överförbarhet, mönstret som kommuniceras känns igen i nya fall. Här resonerar man att överförbarhet kan nås när någon kan förstå olika situationer, processer eller fenomen med hjälp av de tolkningar som gjorts inom forskningen. Svårigheten med detta resonemang är att det bygger på att sammanhanget är baserat på individens tolkningar och ett underliggande antagande kring homogenitet inom ett specifikt sammanhang [6].

Sammanfattningsvis kan man konstatera att den kvalitativa forskningen ofta är beroende av sammanhanget och att läsaren därför noggrant måste fundera över studieresultatens överförbarhet till andra sociokulturella omgivningar. Även om det i studien ska ingå ett resonemang om i vilken utsträckning studieresultaten är överförbara och i linje med tidigare publicerad litteratur, ligger det hos läsaren att bedöma resultatens överförbarhet. Något som underlättar denna bedömning är om man i studien har diskuterat hur resultaten för fram en teoretisk förståelse som är relevant för flera olika situationer. Till exempel kan en studie som undersökt patienters preferenser inom palliativ vård bidra med teorier om etik och humanitet inom hälso- och sjukvården, och på så sätt vara relevant

för andra kliniska sammanhang. Överförbarheten har dock alltid begränsningar. Därför är urvalsstrategin en viktig förutsättning som bidrar till att läsaren kan bestämma var gränserna går för hur överförbara forskningsresultaten är. Detta gäller för all forskning, såväl den som genomförs med kvantitativ metod som den som genomförs med kvalitativ metod [7,8].

8

Allmänt om studier med kvalitativ metod

Även om det finns egenskaper mellan olika kvalitativa forskningsansatser som är identiska, innefattar kvalitativ forskningsmetodik en mängd olika tillvägagångssätt för utförande av studier. Dessa olika ansatser bottnar i olika kunskapsområden inom filosofi (fenomenologi, hermeneutik), antropologi (etnografi) och sociologi (”grounded theory”). Val av metod bestäms av studiens syfte, och den valda metoden tillför på så sätt forskningen ett grundläggande ramverk för formulering av frågeställning, datainsamling, analys och tolkning [9].

Tabell 8.2 Exempel på kvalitativa forskningsansatser. Läs mer om dessa i Bilaga 10.

Ansats	Beskrivning	Exempel på studie
”Grounded theory”	”Grounded theory” används framför allt för att utveckla teorier om människors beteenden genom att analysera kvalitativa data. Metoden innebär att man både formulerar hypoteser utifrån specifik information, och att man drar specifika slutsatser utifrån hypoteser	En studie i två delar om psykologiska effekter av: 1) hel tandlöshet 2) rehabilitering i form av fasta tandersättningar (implantatbroar). Syftet var att kartlägga hur man anpassar livet vid hel tandlöshet, hur man lever med avtagbara proteser och hur fasta tandersättningar påverkar livet i jämförelse med de avtagbara. Författarna lyfter fram tre kategorier: ”att bli en avvikande person”, ”att bli en osäker person”, och efter behandlingen ”att bli den person jag en gång var”. Dessa bildar huvudkategorin ”ändring av självbilden” [10]
Fenomenologi	Fenomenologi är enligt Edmund Husserl (1859–1938) både teori och metod, dvs ett vetenskapsteoretiskt perspektiv och en metodansats (med flera varianter). Fenomenologi handlar om hur vi ger fenomen de betydelser de får, hur de framträder för vårt medvetande och hur våra upplevelser av dessa påverkar vårt sätt att förstå världen (livsvärld)	Sjukgymnaster arbetar med kroppen som bas. Det övergripande syftet med studien var att utveckla en djupare förståelse för hur sjukgymnasten utifrån detta kan hjälpa människor med svårdefinierbara smärt- och stressproblem att återfå sin förlorade hemmastaddhet i sin kropp [11]

Tabellen fortsätter på nästa sida

Tabell 8.2 Fortsättning.

Ansats	Beskrivning	Exempel på studie
Fenomenografi	Fenomenografisk metod har utvecklats inom pedagogisk forskning. Fenomenografi kan definieras som läran om de kvalitativt olika sätt på vilka människor uppfattar aspekter av sin omgivning. Det främsta syftet är att urskilja olika aspekter av fenomenet. Inom fenomenografin är det viktigt att skilja på ”hur något är” och ”hur något uppfattas vara”. Den vanligaste datainsamlingsmetoden är intervjuer	Syftet med studien var att undersöka hur sjuksköterskor i arbetsledande ställning uppfattar munhälsa i allmänhet och vårdtagares munhälsa i synnerhet [11]
Fenomenologisk hermeneutik	Fenomenologisk hermeneutik fokuserar på tolkning av text, t ex intervjuer. Det som tolkas är inte erfarenheter i sig, utan den text som utgörs av de i intervjuerna konstruerade berättelserna. Forskaren strukturerar om texten för att hitta innebörder som ligger under ytan. Delar som har beröringspunkter med varandra förs ihop till större enheter. Intresset riktas mot levda erfarenheter, inte mot personen. Ansatsen bygger på följande metodsteg: narrativa intervjuer, naiv läsning och strukturanalyser	En studie gjordes för att utforska beslutsprocesser vid äggdonationer. Syftet var att undersöka både kvinnors incitament för att donera ägg och deras erfarenheter av att vara potentiella äggdonatorer. Studien använde sig av tolkade intervjuer som datainsamlingsmetod. Intervjumetoden valdes för att uppmuntra kvinnorna att ge uttryck för sina intryck direkt efter första konsultationen, före det att kvinnorna hade tid att processera sina intryck [12]
Etnografi	Det kritiska antagandet som styr etnografisk forskning är att varje grupp av människor som är tillsammans under en tidsperiod kommer att utveckla en kultur. Etnografisk forskning fokuserar på den kultur personer lever i. Den primära metoden inom etnografin är observation (vanligen deltagande observation)	Forskaren studerade patienter som för första gången drabbats av mental sjukdom för att undersöka hur deras livssituation påverkades. Detta gjordes genom att uppleva och identifiera processer inom mentalvården i Köpenhamn [13]
Hermeneutik	Hermeneutik handlar om tolkning och förståelse. I en empirisk studie är det viktigaste analysredskapet just tolkning. Tolkningar presenteras inte som sanningar mellan orsak och verkan, utan som nya och förhoppningsvis givande sätt att förstå känsloreaktioner, motiv för handlingar, tankemönster och andra meningsskapande mänskliga aktiviteter	En hermeneutisk ansats kan vara lämplig för att studera en fråga av existentiell art. I det här exemplet undersökte forskaren hur det är att vara beroende av omvårdnad vid en akutmottagning. Forskaren intervjuade elva patienter och fyra närstående [11]

Tabellen fortsätter på nästa sida

Tabell 8.2 Fortsättning.

Ansats	Beskrivning	Exempel på studie
Kvalitativ innehålls-analys	Innehållsanalys innebär vanligen att forskaren genom upprepad läsning av en text identifierar menings-enheter som sedan kodas. Dessa sorteras sedan i kategorier genom att meningsenheternas likheter och skillnader jämförs. Inget material får exkluderas för att det saknas lämplig kategori. Inget material får heller falla mellan två kategorier, eller passa in i mer än en kategori	I en studie vars syfte var att belysa upplevelser av ensamhet bland de allra äldsta intervjuades 30 personer mellan 85 och 103 år. Eftersom upplevelsen av ensamhet kan variera från individ till individ, och eftersom kvalitativ innehålls-analys används för att identifiera likheter och skillnader i en text ansågs ansatsen lämplig för ändamålet [11]
Aktions-forskning	Aktionsforskning syftar till att lösa specifika problem inom ett program, en organisation eller ett samhälle. Generellt kopplas aktionsforskning samman med handlingar som leder till förändring och utveckling. Det mest påfallande kännetecknet för aktions-forskning är ett deltagarbaserat och interaktivt förhållningssätt. Detta innebär att alla deltagare, både forskare och praktiskt verksamma personer arbetar tillsammans	Syftet med studien var att arbeta fram och att införa en lämplig modell för handledning av vårdpersonal inom kriminalvården [13]
Narrativ metod	Narrativ metod är lämplig att använda om man vill öka kunskapen kring mening och mönster i personers berättelser om sig själva och sina liv. Det finns inte en enskild narrativ analysmetod, utan flera komplette-rande metoder som alla bygger på den filosofiska och teoretiska grundvalen att mänsklig förståelse har en narrativ form	I en studie som syftade till att undersöka vägen ur missbruk och hemlöshet ur ett aktivitetsperspektiv samlades data in med hjälp av narrativa intervjuer med före detta hemlösa kvinnor. Analysens inriktning var att först kategorisera för att sedan tolka utifrån narrativa utgångspunkter såsom mening eller mönster [11]

Urval

Målet är att göra ett urval som leder till en ökad *förståelse* för variationer i det fenomen som studeras. Inom ”grounded theory” används *teoretiskt urval* (”theoretical sampling”). Det betyder att man inte i förväg bestämmer vilka personer och hur många informanter som ska ingå i studien. Man börjar med få deltagare och efterhand som analysen fort-skrider lägger man till fler för att induktivt kunna skapa teori om det studerade fenome-net. Vanligen får man då en heterogen grupp av informanter.

En annan metod för urval är *strategiskt urval* (”purposeful sampling”). Denna urvalsme-tod innefattar olika tillvägagångssätt som lämpar sig för olika sorters studier beroende på syfte och förutsättningar. Man kan exempelvis använda sig av följande urvalsmetoder [13,14].

Snöbollsurval (”Snowball” eller ”chain sampling”, sociogram)
En metod som används för att hitta informationsrika personer eller kritiska fall. Genom att fråga ett visst antal personer om vem annan man bör tala med ökar snöbollen i omfång när nya informanter hittas. De personer som rekommenderas som värdefulla av flera informanter är speciellt viktiga.

Maximal variation (”Maximum variation sampling”)
Syftet är att fånga och beskriva fenomenets variation i olika sammanhang. Om urvalet är för litet och det råder stor heterogenitet kan problem uppstå vad gäller förståelse av fenomenet eftersom individuella fall är så olika varandra. Istället fokuserar man på att hitta information som belyser variationen av fenomenet och signifikanta gemensamma mönster inom det varierade urvalet.

Avvikande fall (”Extreme or deviant case sampling”)
Fokuserar på personer som är rika på information eftersom de är ovanliga eller speciella på något sätt.

Homogent urval (”Homogenous samples”)
Syftet med denna metod är att beskriva en specifik undergrupp på djupet. Fokusgruppintervjuer är ofta baserade på ett homogent urval. Då samlas personer med liknande bakgrund och erfarenheter för att delta i fokusgruppintervjuer om specifika ämnen som berör dem.

Bekvämlighetsurval (”Convenience sampling”)
Den minst önskvärda urvalsmetoden. Forskaren tänker att eftersom urvalet inte är stort nog för att producera överförbara slutsatser kan han välja informanter som är lättillgängliga och billiga att studera. Bekvämlighetsurval är varken en meningsfull eller strategisk urvalsmetod.

Datainsamlingsmetoder
Inom kvalitativ forskning kan man använda sig av olika datainsamlingsmetoder (Tabell 8.3). Vilken metod som väljs är beroende av vad man ämnar studera. Om man vill studera upplevelser (erfarenheter, åsikter, känslor, behov och önskemål) kan intervjuer vara en lämplig metod. Observation är lämpligare för studier av beteenden (interaktion mellan personer, grupprocesser, könsrollsmönster osv) [13,14].

Intervjuer kan ha olika struktur, t ex öppen, semistrukturerad eller helt strukturerad. Intervjuer kan även utföras på olika sätt, t ex djupintervju av en enskild person, intervju av personer på gatan, eller fokusgruppintervjuer.

Observationer kan vara deltagande eller icke-deltagande. Personer man observerar kan ha olika grad av medvetenhet kring att en observation sker och vad som observeras. Observationer dokumenteras med hjälp av anteckningar eller videoupptagning.

Enkätundersökningar med öppna frågor.

Tabell 8.3 Exempel på kvalitativa datainsamlingsmetoder [2].

	Djupintervjuer	**Fokusgrupper**
Kännetecken	Individuell intervju i vilken deltagaren uppmuntras att berätta i detalj om sitt perspektiv av forskningsämnet	• Gruppdiskussion som leds av en forskare/moderator • Gruppen består av ca 6–10 personer • Grupperna kan vara naturligt förekommande eller bestå av rekryterade deltagare • Interaktionen mellan deltagarna genererar data
Typ av data	Genereras	Genereras
Form av data	Utskrift av intervju (ordagrann)	Utskrift av gruppdiskussion (ordagrann)
När använda	• När deltagarens uppfattning är viktig • När komplexa processer eller erfarenheter studeras • När övertygelser, förståelse och tolkningar av ett fenomen studeras • När besluts- och motivationsprocesser studeras • När konfidentiella ämnen studeras • Användbar om urvalet är utspritt geografiskt	• När deltagarnas uppfattning är viktiga • När man undersöker hur attityder, kunskap och övertygelser uppstår och hur de utmanas • När man undersöker sociala normer • När kreativt tänkande krävs • När deltagare känner att de inte har något att säga • När deltagare kan känna sig stärkta av att diskutera känsliga ämnen med dem som upplevt samma sak
Val av datainsamlingsmetod	Intervjuer är en bra metod om man vill fånga in deltagarnas perspektiv gällande ett fenomen, eller för att utforska övertygelser, motivations- eller beslutsprocesser Djupintervjuer är bra om man vill undersöka komplexa processer eller erfarenheter	• Man bör vara försiktig om det råder obalans i makt eller status mellan deltagarna • Man bör vara försiktig om känsliga ämnen ska diskuteras

Tabellen sträcker sig över hela uppslaget

Skriftligt material, t ex dagböcker, protokoll, berättelser, journaler och litteratur.

Inom de olika kvalitativa forskningsansatserna kan man använda samma metoder för datainsamling. Skillnaden ligger då i hur man enligt forskningsansatsens metodik analyserar och tolkar materialet.

	Observation	**Dokumentanalys**
	Forskaren observerar beteenden, skeenden, och interaktioner. Inkluderar: • Deltagande observation (forskaren deltar jämte studiedeltagarna i studien för att uppleva fenomenet själv) • Direkt observation (forskaren observerar studiedeltagarna men behåller sin oberoende ställning)	Forskaren studerar dokument för att utforska dess innehåll och mening
	Naturligt förekommande	Naturligt förekommande
	• Videoinspelningar • Bandinspelningar • Fältanteckningar	Dokument såsom officiella rapporter, minnesanteckningar, ledarsidor, dagböcker, brev, tidskrifter, affischer etc
	• När naturlig kontext är mycket viktigt • Vid jämförelse av faktiskt och rapporterat beteende • När kommunikation, social ställning, beteenden, eller interaktioner utforskas • När implicita aspekter, och sådant som tas för givet, avseende beteendet ska undersökas	• Användbart för att utforska dominerande diskurser, rådande normer, förklaringar av fenomen • Vid jämförelse av offentliga och privata värderingar • Historisk forskning
	Observationsmetoder är bra om olika aspekter av forskningsämnet är omedvetna, tagna för givet, eller om deltagaren inte är villig att diskutera ämnet uppriktigt	–

Urvalsstorlek

Det finns inga regler för hur stort urvalet ska vara inom kvalitativ forskningsmetodik, utan urvalet bestäms generellt av informationsbehovet. En guidande princip vid datainsamling är datamättnad, vilket innebär att mängden insamlad data som krävs för en specifik studie skiftar beroende på hur snabbt forskaren anser att man har kommit till den fas då ytterligare datainsamling inte ger mer kunskap – alltså när man har nått mättnad. Det är ändå vanligt att man fortsätter med ytterligare lite datainsamling för säkerhets skull. Antalet informanter som behövs för att mättnad ska nås beror exempelvis på hur avgränsad frågeställningen är. En bredare frågeställning kräver sannolikt fler informanter innan mättnad nås än vad en avgränsad frågeställning gör. Annat som kan påverka är hur bra informanterna kan reflektera över exempelvis sina erfarenheter och hur de kommunicerar dessa och hur duktig forskaren är på att samla in data från informanterna eller observationerna. Begreppet mättnad kommer från ”grounded theory”, men det används även inom andra kvalitativa ansatser [9].

Analys

Eftersom forskaren ofta utför datainsamlingen blir denne också en del av analysen som påbörjas redan under datainsamlingens gång. Det finns olika sätt att analysera kvalitativa data. Tillvägagångssättet bestäms ofta av det teoretiska perspektiv eller den forskningsansats som studien är baserad på. Läs mer om olika forskningsansatser och analysmetoder i Bilaga 10.

Utvärdering och syntes av studier med kvalitativ metod

Formulera frågeställning

När man gör en översikt och ska söka efter kvantitativa studier brukar man formulera sin frågeställning och sökning enligt PICO-modellen, där P står för population, I för intervention, C för kontroll och O för ”outcome” (effektmått) (Kapitel 3). När man formulerar frågeställningar för studier utförda med kvalitativ metod kan SPICE-modellen vara mer användbar. S står för ”setting” (sammanhang), P för ”perspective” (perspektiv), I för ”intervention/interest” (intervention), C för ”comparison” (jämförelse) och E för ”evaluation” (utvärdering) [15,16].

Modellen är i flera avseenden överensstämmande med PICO, dvs P i PICO motsvaras av S och P i SPICE och I i PICO motsvaras av C i SPICE. Slutligen har O i PICO likheter med E i SPICE. Alla komponenter är nödvändigtvis inte representerade i varje studie och modellen ska mer ses som en underlättande guide för strukturering av frågeställning och litteratursökning.

Tabell 8.4 SPICE-modell för formulering av frågeställning.

”Setting” (sammanhang)	**”Perspective” (perspektiv)**	**”Intervention/ Interest” (intervention)**	**”Comparison” (jämförelse)**	**”Evaluation” (utvärdering)**
Var? Kontexten i studien, exempelvis en kultur, ett sjukvårdssystem eller ett område	För vem? Det perspektiv som uppvisas genom olika värderingar eller attityder	Vad? Det fenomen som studeras	Något annat? Jämförelse (alla studier har inte en jämförande komponent)	Vilket resultat? Utvärdering som innefattar både process och resultatutvärdering

Exempel 8.1 Frågeställning enligt SPICE.

Frågeställning: Hur upplever föräldrar till barn med cancer sin livskvalitet i hemmet?

S	**P**	**I**	**C**	**E**
Hemmet	Föräldrar till barn med cancer	Cancer hos barn	Inte tillämpbart	Upplevelse av livskvalitet

Litteratursökning

Att skapa en sökstrategi för att identifiera studier med kvalitativ metod följer i väsentliga delar den process som beskrivs i Kapitel 4, ”Litteratursökning”. Sökningarna ska alltså vara systematiskt uppbyggda och kunna reproduceras.

En viktig skillnad att uppmärksamma när man utformar sökstrategier för att identifiera studier baserade på kvalitativ forskningsmetodik jämfört med kvantitativ, är att det finns brister i databasernas indexering och/eller i artiklarnas abstrakt. En central databas som Medline (PubMed) erbjuder inte någon differentierad indexering av olika forskningsstrategier som försöker fånga individens subjektiva upplevelse, med utgångspunkt från hur individen själv uttrycker denna. Andra hinder är att själva utformandet av rubriker och abstrakt kan leda till svårigheter i sökprocessen. Viktig information kan saknas i de fall då författarna valt att inte strukturera abstrakten enligt modellen syfte-metod-resultat-slutsats-ämnesord. Sökning på ord i titlar försvåras av att dessa kan sakna direkt koppling till det som artikeln handlar om. Avsaknaden av den här typen av information kan i sin tur leda till att artikeln inte indexeras med något kontrollerat ämnesord (deskriptor) som relaterar till det som vanligen uttrycks som ”kvalitativ forskning” eller ”kvalitativa studier”. På grund av problemen med indexering är det svårt att utforma sökningar som identifierar specifika kvalitativa forskningsmetoder.

Exempel på andra söksätt för att finna studier med kvalitativ forskningsmetodik är att använda *Related articles* i PubMed, genom att granska referenslistor eller genom att citeringssöka [17–19].

Val av databas

Frågeställningen styr alltid valet av databaser, oavsett vilken typ av studier som är i fokus. I kapitlet om litteratursökning (Kapitel 4) beskrivs några centrala databaser: PubMed, Embase, Cochrane Library, CINAHL och PsycINFO. Ibland kan man även behöva söka i kompletterande databaser. Exempel på sådana är Sociological Abstracts, alternativt SocIndex och Social Services Abstracts som är breda databaser inom sociologi och socialvetenskap. Ett annat skäl att söka i dessa ämnesdatabaser är att olika databaser har olika indexeringsprinciper. Mer generella databaser såsom Academic Search Elite, citeringsdatabaserna i Web of Science, alltså även Social Sciences Citation Index, kan av samma skäl ge andra utgångspunkter för sökningen samtidigt som man kan få träff på studier som inte är registrerade i de stora biomedicinska databaserna. Ytterligare en typ av stöd i sökprocessen ger plattformar såsom Elsevier's SciVerse. Innehållet i plattformen är beroende av vilka databasprenumerationer som finns. Databaser som återfinns här kan vara ScienceDirect, citeringsdatabasen Scopus och även den fria PubMed, som inte är en Elsevier-databas. SciVerse samsöker och räknar upp antal träffar i respektive databas vilket ger en översiktlig bild av vilka databaser som bäst matchar olika sökord och söksträngar. Plattformen kan därför vara en mycket bra utgångspunkt för testsökningar i allmänhet, och för mer svårsökta studier i synnerhet. Den kan även ge unika ämnesträffar.

Utgå från redan existerande översikter eller syntheser

När ett nytt sökprojekt startar börjar man med att kontrollera om det redan finns liknande systematiska översikter på ämnet. Redan existerande översikter har betydelse för projektplaneringen och därmed också för litteratursökningen. Översikter kan också vara bra vid planering och utformning av sökstrategier. Andrew Booth diskuterar och redovisar i "Supplementary Guidance for Inclusion of Qualitative Research in Cochrane Systematic Reviews of Interventions" flera olika redan utarbetade sökstrategier för systematiska översikter, oavsett design. Han har också utarbetat en söksträng för att specifikt söka det han kallar "qualitative systematic reviews", dvs systematiska översikter där kvalitativa studier har syntetiserats [20]:

> qualitative systematic review* OR (systematic review AND qualitative) OR evidence synthesis OR realist synthesis OR (qualitative AND synthesis) OR meta-synthesis* OR meta synthesis* OR metasynthesis OR meta-ethnograph* OR metaethnograph* OR meta ethnograph* OR meta-study OR metastudy OR meta study

Sökfilter

Sökfilter (på engelska ”search filters” eller ”hedges”) är ett hjälpmedel för att underlätta sökningen av en viss studiedesign, eller särskilda aspekter som biverkningar eller diagnoser. Filter är en söksträng vars känslighet och träffsäkerhet har värderats. Olika filter skapas och anpassas också till olika databaser. Syftet är bl a att effektivisera informationssökningen. Söksträngarna består vanligen av både kontrollerade sökord ur databasernas tesaurus[1] och av fritextord, dvs vanligt förekommande ord i databasernas beskrivning av ingående studier från bl a artiklars titlar och abstrakt. Sökfilter kombineras sedan med sökord för det aktuella ämnet.

Det finns flera organisationer som arbetar med att utforma och värdera sökfilter, däribland också filter för sökning av studier som baseras på kvalitativa data. The Hedges Project vid McMaster University i Kanada har utvecklat sökfilter för bl a Medline, Embase, PsycINFO och CINAHL. Vid Centre for Reviews and Dissemination (CRD) i Storbritannien arbetar en grupp informationsspecialister med att identifiera och värdera redan skapade filter. Dessa filter är samlade på adressen www.york.ac.uk/inst/crd/intertasc/qualitat.htm.

Söka med kontrollerade sökord

Trots att det finns ett antal sökfilter för ”qualitative studies” kan det av olika skäl ändå vara så att man väljer att skapa en söksträng. Det kan handla om att filtren inte är anpassade till aktuellt ämnesområde eller databas, eller att det är föråldrat i förhållande till förändringar som gjorts i en databas innehåll samt indexering.

Databaser som täcker forskningsområden, där studier baserade på kvalitativa data är vanliga, erbjuder, som tidigare nämnts, ofta en mer indelad indexering. CINAHL har länge haft flera användbara kontrollerade sökord. Förutom det allmänna *Qualitative Studies,* finns även indexeringsord som t ex *Ethnographic Research, Ethnological Research, Phenomenological Research* och *Grounded Theory*. I Medline finns betydligt färre kontrollerade ämnesord. Det mer övergripande ämnesordet *Qualitative Research* infördes i MeSH-databasen först 2003. Därutöver finns t ex *Nursing Methodology Research* och *Focus Groups,* som har funnits med i MeSH-databasen sedan 1990-talet.

En tesaurus är ett hjälpmedel som både förändras och utvecklas över tid. Nya termer tillkommer när nya aspekter av olika ämnesområden utvecklas, eller när nya begrepp börjar användas i forskningen. Äldre indexeringsord tas bort eller ersätts. För att systematiskt identifiera vilka olika indexeringsord som finns i de olika databaserna använder man data-

[1] En tesaurus är en förteckning över gruppvis likvärdiga eller synonyma ord och fraser, så kallade deskriptorer eller nyckelord som används för att karakterisera dokument i en databas.

basspecifika tesaurusar. I CINAHL with Full Text via EBSCO ligger tesaurusen under rubriken *Cinahl Headings*. Där är *Research Methodology* en användbar ingång för att få en bättre överblick. I Medline (PubMed) är det bra att börja med *MeSH browser* eller *MeSH database*. Där kan man välja det övergripande *Research* eller några steg ned i hierarkin *Qualitative Research*, och termen *Nursing Research*. I PsycINFO via EBSCO under rubriken *Thesaurus* kan *Methodology* vara en bra start för att identifiera indexeringsord.

8

Söka med fritextord

Precis som vid annan litteratursökning bör man vid en litteratursökning av studier utförda med kvalitativ forskningsmetodik kombinera tesaurustermer med fritexttermer för att fånga upp icke-indexerade studier. Ofta är en avgränsning till fälten titel och abstrakt i databaserna att föredra för att minska åtminstone en del av det brus som en fritextsökning kan föra med sig. Att söka med fritexttermer är extra viktigt i databaser med mindre urval av tesaurustermer så man kan fånga upp olika teoretiska ansatser som är vanliga inom den här typen av forskning.

Förslag till söksträng i PubMed

Här ges förslag på sökord som kan användas vid utformning av olika sökstrategier:

> ”Qualitative Research” [MeSH] OR ”Focus Groups”[MeSH] OR ”Nursing Methodology Research”[MeSH]” OR qualitative[Title/Abstract][2] OR ”grounded theory”[Title/Abstract] OR ethnogra*[Title/Abstract] OR ethnolog*[Title/Abstract] OR phenomenogra*[Title/Abstract] OR phenomenolog*[Title/Abstract] OR hermeneutic*[Title/Abstract] OR focus group*[Title/Abstract] OR field study[Title/Abstract] OR narrativ*[Title/Abstract] OR lived experience* [Title/Abstract] OR life experience*[Title/Abstract]

Urval av litteratur

Två personer (eller fler) granskar, oberoende av varandra, abstraktlistor från databassökningarna. Studier som bedöms vara relevanta utifrån rubrik och abstrakt tas fram i fulltext. Det räcker att en av granskarna anser att artikeln bör läsas i fulltext för att den ska tas fram. Granskarna bedömer därefter, oberoende av varandra, om de beställda artiklarna är relevanta för frågeställningen samt uppfyller eventuella andra inklusionskriterier.

[2] Ett uttryck som ”qualitative” kan ge alltför många träffar då termen används i många betydelser. I vissa sökningar behöver man specificera:

qualitative study[Title/Abstract] OR qualitative research[Title/Abstract] OR qualitative descriptive study[Title/Abstract] OR qualitative data[Title/Abstract] OR qualitative content analysis[Title/Abstract] OR qualitative interview*[Title/Abstract] OR qualitative approach[Title/Abstract] OR qualitative analysis[Title/Abstract] OR qualitative design[Title/Abstract] OR qualitative exploration[Title/Abstract] osv.

De artiklar som inte uppfyller kriterierna sorteras bort. Granskarna jämför därefter sina inklusionslistor. Om listorna inte överensstämmer, diskuterar granskarna inbördes och beslutar huruvida artikeln ska inkluderas eller inte.

Kvalitetsbedömning

De kriterier som används för att bedöma den vetenskapliga tillförlitligheten i kvalitativa studier är i mångt och mycket desamma som används för kvantitativa studier. Studien ska ha hög läsförståelse och vara logiskt strukturerad. Det bör framgå varför forskaren valt att använda sig av kvalitativa metod(er) för att generera och/eller analysera data, samt varför man valt en specifik kvalitativ ansats. Frågeställningar ska vara väldefinierade. Urval och kontext bör vara relevant och tydligt beskrivna. Metodavsnittet ska vara grundligt redovisat, så att läsaren kan bedöma om datainsamlingen och analysmetoden verkar vara adekvat. Forskaren bör även redovisa hur data och resultat relaterar till varandra, hur analysprocessen gått till och om det finns någon teorianknytning. Resultat och tolkningar ska beskrivas logiskt och begripligt. Forskaren bör även argumentera kring resultatens överförbarhet i relation till både urval och kontext. Den vetenskapliga kvaliteten kan bedömas som högre om pålitligheten och giltigheten hos data är hög och om analysproblem diskuteras, alltså om tolkningen har verifierats och problem har hanterats på lämpligt sätt [21–23]. För att ytterligare underlätta granskningsprocessen vid kvalitetsbedömningen kan man använda sig av en granskningsmall (Bilaga 5).

När det gäller forskarens roll i studien finns det inom kvantitativ forskningstradition ett strävande efter distans och objektivitet, medan forskaren ofta är betydligt mer involverad i kvalitativa studier. Ofta är forskaren själv ett redskap för urval, datainsamling och analys och därav blir forskarens roll och förståelse en viktig komponent i kvalitetsbedömningen av studien. Med förförståelse menas det ”bagage” som forskaren bär med sig in i ett forskningsprojekt. Förförståelsen påverkar forskaren under hela projektets gång, exempelvis under datainsamling och dataanalys. Förförståelsen innefattar forskarens hypoteser, erfarenheter, yrkesmässiga perspektiv och den teoretiska referensram som forskaren har när projektet påbörjas. Generellt är förförståelsen en viktig del av forskarens motivation för att inleda forskning kring ett särskilt ämne, men den kan även bidra till att forskaren går in i ett projekt med begränsad öppenhet och begränsad förmåga att lära sig av datamaterialet. Forskaren ska eftersträva att ha ett aktivt och medvetet förhållande till sin förförståelse. Forskaren bör därför i metoddiskussionen redogöra för hur denne hanterat sin förförståelse samt redogöra för sin roll i studien [8].

En sak som kan orsaka förvirring gällande beskrivning av kvalitativ forskning är användningen av olika begrepp. I en del vetenskapliga artiklar används begrepp som är relaterade till kvantitativ tradition, exempelvis trovärdighet. För att bedöma trovärdigheten hos kvantitativa resultat granskar man ofta validiteten, reliabiliteten och generaliserbarheten

hos resultaten. Inom kvalitativ tradition används ofta andra begrepp. Giltighet (”credibility”) avser datainsamlingens och analysens trovärdighet. Med tillförlitlighet (”dependability”) avses om forskningen är oberoende forskaren och dennes perspektiv. Det är också viktigt att bedöma om andra kan bekräfta det som forskaren fått fram (”confirmability”) samt bedöma forskningens överförbarhet (”transferability”). Det finns idag olika åsikter kring vilka begrepp som bör användas och därför är det viktigt att känna till båda traditionernas begrepp [8,24]. Se Tabell 8.5 för bedömning av studiens vetenskapliga kvalitet.

Tabell 8.5 Kriterier för bedömning av vetenskaplig kvalitet [25].

Hög kvalitet	Medelhög kvalitet	Låg kvalitet
Klart beskrivet sammanhang (kontext)	Sammanhanget ej beskrivet tydligt (kontext)	Oklart beskrivet sammanhang (kontext)
Väldefinierad frågeställning	Frågeställning ej beskriven tydligt	Vagt definierad frågeställning
Välbeskriven urvalsprocess, datainsamlingsmetod, transskriberingsprocess och analysmetod	Några otydligheter i beskrivningen av urvalsprocess, datainsamlingsmetod, transskriberingsprocess och analysmetod	Otydligt beskriven urvalsprocess, datainsamlingsmetod, transskriberingsprocess och analysmetod
Dokumenterad metodisk medvetenhet	Några otydligheter i den dokumenterade metodiska medvetenheten	Dåligt dokumenterad metodisk medvetenhet
Systematisk, stringent presentation av data	Otydligheter i presentationen av data	Osystematisk och mindre stringent dataredovisning
Tolkningars förankring i data påvisad	Några otydligheter om tolkningars förankring i data	Otydlig förankring av tolkningarna i data
Diskussion om tolkningarnas trovärdighet och tillförlitlighet	Några otydligheter om tolkningarnas trovärdighet och tillförlitlighet	Diskussion om tolkningarnas trovärdighet och tillförlitlighet är bristfällig eller saknas
Kontextualisering av resultat i tidigare forskning	Otydlig kontextualisering av resultat i tidigare forskning	Kontextualisering av resultat i tidigare forskning saknas eller är outvecklad
Implikationer för relevant praktik välformulerade	Implikationer för relevant praktik är otydligt beskrivna	Implikationer för relevant praktik saknas eller är otydliga

Syntes

Efter granskningsprocessen tabelleras studierna (Tabell 8.2.4) och grupperas sedan enligt metod/forskningsdesign. I de fall där råmaterialet är text som avspeglar det informanterna berättat ska kategoriseringen kunna visas med citat. Om detta inte är fallet, exempelvis vid observationsstudier eller vid aktionsforskning, ska det framgå hur kategorierna i syntesen har bildats, eller om de är baserade på de ursprungliga kategorierna i de inkluderade studierna. Vid syntes av resultat från studier baserade på olika forsknings-

ansatser bör man vara aktsam och problematisera kring vilka val man gjort. Därefter inleds syntesen, vilket innebär att resultaten från studierna kombineras för att skapa ett nytt perspektiv eller synsätt. Syntesprocessen kan t ex utföras enligt Howell Major och Savin-Baden [26]. På samma sätt som vid urvalet av studier görs syntesen av två (eller flera personer) där man först gör en oberoende syntes som sedan diskuteras i par (grupp) för att nå konsensus.

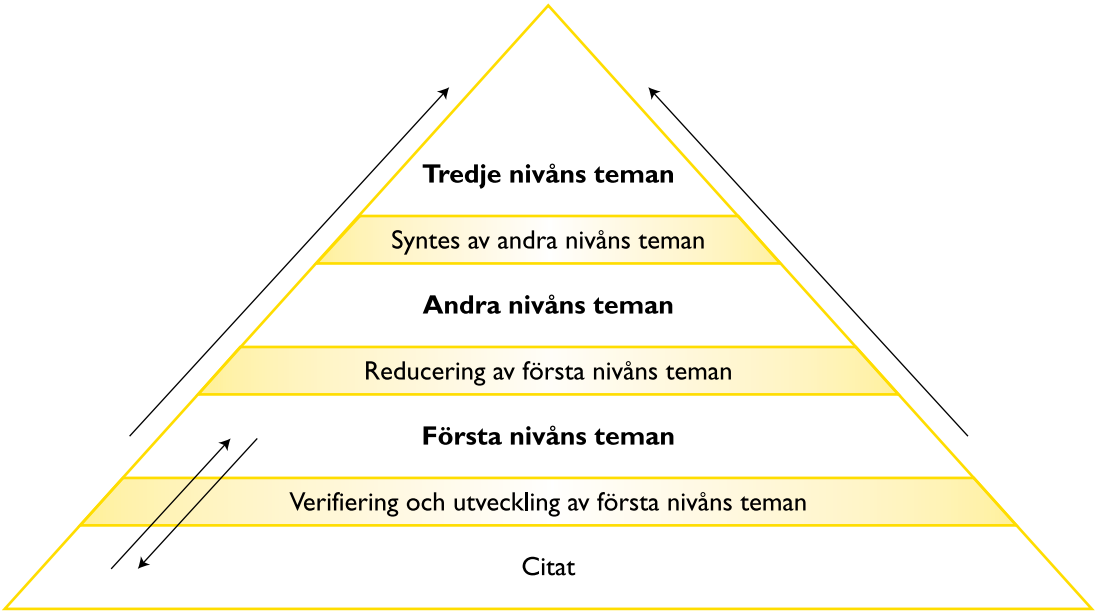
Syntesen görs sedan i fyra steg (Figur 8.1):

1. De kvalitetsgranskade studierna som inkluderats gås igenom för att identifiera resultat i form av teman (koder, kategorier eller subkategorier). Dessa teman bör sedan stämmas av mot eventuella citat. Därefter undersöker man om något tema framkommer i flera studier. Dessa kondenseras sedan under utvecklandet av *första nivåns teman.*

2. Besläktade *första nivåns teman* reduceras sedan till *andra nivåns teman.* Detta är en komplex och dynamisk process i vilken de olika temana arrangeras om i flera steg till dess att tydliga andra nivåns teman framkommer.

3. Besläktade *andra nivåns teman* syntetiseras slutligen till övergripande *tredje nivåns teman*. Viktiga mönster och samband bland andra nivåns teman tolkas och problematiseras. Processen repeteras tills tredje nivåns teman fastställs.

4. En samlad bedömning görs av det vetenskapliga underlaget varefter evidensgraderade resultat och slutsatser formuleras [26,27].



Figur 8.1 Syntesprocessen för studier med kvalitativ analys.

Det går inte att sammanväga resultaten enligt GRADE (Kapitel 10). Analysprocessen har dock likheter med GRADE, där de enskilda temana skulle kunna ses som effektmått. Liksom resultaten per effektmått vägs samman i GRADE, så vägs även resultaten per tema samman i den kvalitativa syntesen.

Evidensstyrkan i resultaten bedöms enligt följande:
Det finns vetenskapligt stöd – Identifierade studier har tillräcklig kvalitet och relevans.

Det vetenskapliga underlaget är otillräckligt – Identifierade studier saknar tillräcklig kvalitet och relevans.

Bedömning av evidensstyrkan göras av minst två personer under konsensusförfarande.

Exempel 8.2 Syntes om upplevelse av tandlöshet.

Syntesen i sin helhet är redovisad i SBU-rapporten *Tandförluster* [28].

Frågeställning: Hur upplever människor att förlora tänder och att vara tandlös?

1. De inkluderade studierna gås igenom för att identifiera resultat i form av teman (koder, kategorier eller subkategorier). Dessa teman stäms sedan av mot eventuella citat. Därefter undersöker man om något tema framkommer i flera studier. Dessa kondenseras sedan under utvecklandet av *första nivåns teman*.

Tabell 8.2.1 visar hur första nivåns tema ”sorg och skam” kan uttryckas som citat.

Tabell 8.2.1 Exempel på hur första nivåns tema ”sorg och skam” har uttryckts i citaten och reducerats och syntetiserats till andra och tredje nivåns teman.

Citat med specifikt fokus	Strukturerat citat	Kondenserat citat	Syntes	Första nivåns tema	Andra nivåns tema	Tredje nivåns tema
”I’ve found when I’m speaking to people I tend to be looking at their teeth and thinking. What lovely teeth you’ve got. Silly. I know. I didn’t do that before… Here’s me with these horrible false teeth”	I säger att när hon pratar med människor så har hon tendens att titta på deras tänder och tänka: Vilka vackra tänder du har. Så gjorde I inte förut… Här är I med sina förfärliga konstgjorda tänder	I brukar numera, när hon pratar med andra, uppmärksamma om de har vackra tänder. Det, i motsatts till I:s egna fula konstgjorda tänder, proteser	I noterar om andra har vackra tänder och jämför med sina egna fula proteser	Sorg och skam	Sänkt självkänsla	Förlust av livskvalitet

I = Informanten

Exemplet fortsätter på nästa sida

Exempel 8.2 Fortsättning.

2. Besläktade *första nivåns teman* reduceras sedan till *andra nivåns teman*. Detta är en komplex och dynamisk process i vilken de olika temana arrangeras om i flera steg till dess att tydliga andra nivåns teman framkommer, dvs flera första nivåns teman reduceras till färre andra nivåns teman.

Nedan följer exempel på *andra nivåns tema* ”sänkt självkänsla”.

Sänkt självkänsla omfattar tre första nivåns teman: inverkan på självförtroende, utseende och sorg/skam. Tandlöshet upplevs som att man är mindervärdig som människa och man känner sig amputerad. Det sviktande självförtroendet kan också påverka könsrollen.

> *”I’ve always been quite fit and that person, suddenly to find that person, that part of me, is going downhill.”*

> *”One feels somewhat misplaced, handicapped. Not human in a way.”*

Man lever med ständig inre osäkerhet och är rädd för att tandlöshet eller dåliga tänder ska uppfattas som något komiskt. Detta kan beskrivas som en sorts oral överkänslighet.

> *”If someone laughed, I thought they were laughing at me.”*

> *”I feel embarrassed to go to bed, you turn your back because I feel my partner will keep laughing, and so… your confidence is gone.”*

Det förändrade utseendet känns som ett för tidigt åldrande, man sörjer sin förlorade ungdom och det kan vara så outhärdligt att man inte vill se sig i spegeln.

> *”Because a person with no teeth looks older. There is no way to say that they don’t because they do, you know?”*

Många upplever sorg över sina förlorade tänder och skäms för att inte ha en munhälsa som är normal för andra.

> *”I’ve found when I’m speaking to people I tend to be looking at their teeth and thinking. What lovely teeth you’ve got. Silly. I know. I didn’t do that before… Here’s me with these horrible false teeth.”*

I dagens samhälle är det avvikande att ha synliga tandluckor eller dåligt fungerande avtagbara proteser. Munnen och tänderna upplevs som en känslig och privat sfär som man inte diskuterar, och man vill bokstavligt talat inte tappa ansiktet ens för sina allra närmaste genom att visa sig utan tänder.

Exemplet fortsätter på nästa sida

Exempel 8.2 Fortsättning.

3. Besläktade *andra nivåns teman* syntetiseras slutligen till övergripande *tredje nivåns teman*. Viktiga mönster och samband bland andra nivåns teman tolkas och problematiseras. Processen repeteras tills tredje nivåns teman fastställs. Studier som ligger till grund för syntesen redovisas i syntestabell och gärna som en exempelpyramid.

Se Tabell 8.2.2 för exempel på syntestabell för *tredje nivåns tema* ”förlust av livskvalitet”.

Tabell 8.2.2 Tredje nivåns tema: förlust av livskvalitet.

Andra nivåns tema	Första nivåns tema	Studier
Sänkt självkänsla	Inverkan på självförtroendet	De Palma, De Souza e Silva, Fiske, Smith, Trulsson*
	Utseende	De Palma, De Souza e Silva, Fiske, Smith, Trulsson*
	Sorg/Skam	Fiske, Smith
Lägre social status	Socialt stigma	De Palma, Fiske, Smith, Trulsson*
	Social kompetens	De Palma, Fiske, Smith, Trulsson*
	Demaskeringsångest	Fiske, Graham, Trulsson*
	Tabu	Fiske
Försämrad funktion	Funktionsförlust	De Palma, Fiske, Smith
	Funktionshinder av smärta	Trulsson*
Förlusthantering	Anpassning	De Palma, Fiske, Graham, Trulsson*
	Självförebråelse	De Palma, Fiske, Graham, Smith
	Bortförklaring	Trulsson*

** Studie med hög kvalitet. Övriga studier är av medelhög kvalitet.*

4. De evidensgraderade resultaten sammanställs och formuleras i form av sammanfattande löptext och resultattabell.

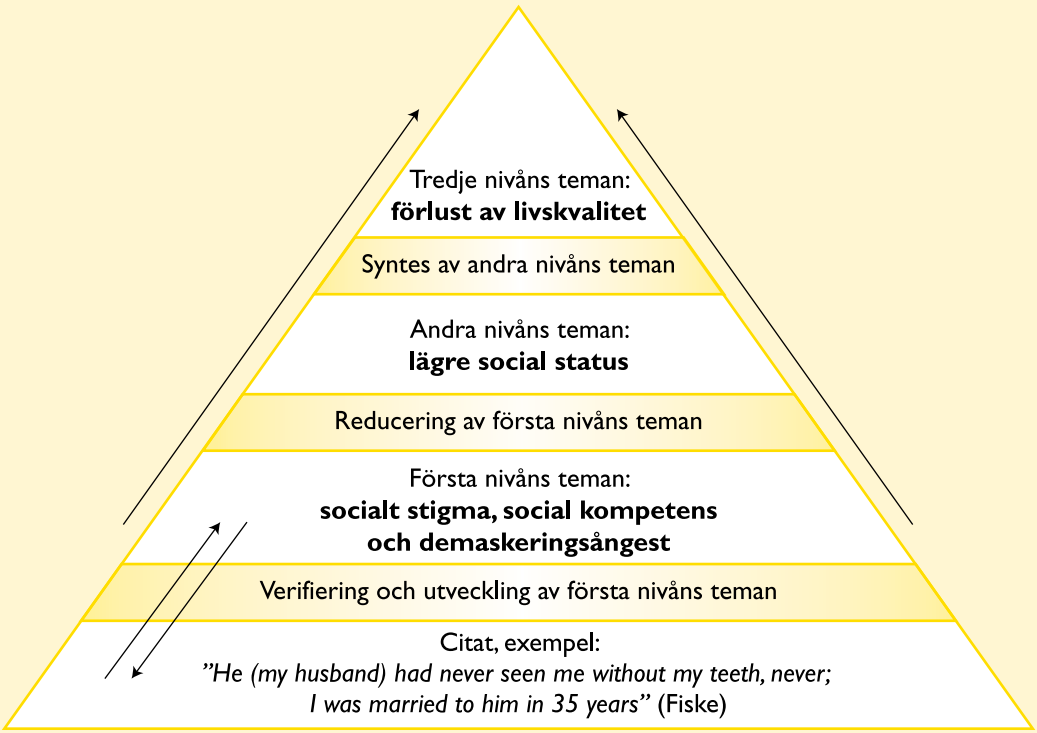
Nedan beskrivs exemplet ”sänkt självkänsla”. Den sammanfattande inledande löptexten avser flera *andra nivåns teman.* Tabell 8.2.3 visar *andra nivåns tema* ”sänkt självkänsla”.

Tredje nivåns tema: förlust av livskvalitet

- Det finns vetenskapligt stöd för att personer som förlorar tänder upplever sänkt självkänsla, lägre social status och försämrad funktion. Personen kan möta denna förlust på olika sätt.

Exemplet fortsätter på nästa sida

Exempel 8.2 Fortsättning.



Figur 8.2.1 Exempel på hur de olika temanivåerna utvecklades. Verifiering och utveckling av första nivåns teman resulterade i ”socialt stigma”, ”social kompetens” samt ”demaskeringsångest”. Dessa reducerades sedan till andra nivåns tema, ”lägre social status”. Slutligen syntetiserades ”lägre social status” tillsammans med de övriga andra nivåns teman till det övergripande tredje nivåns tema ”förlust av livskvalitet”.

Tabell 8.2.3 Andra nivåns tema: sänkt självkänsla.

Andra nivåns tema: *sänkt självkänsla*	**Länder Studiekvalitet Antal studier (antal informanter)**	**Beskrivning**
Första nivåns teman *Självförtroende, utseende och sorg/skam*	Brasilien, England, Sverige 1 hög kvalitet, 4 medelhög kvalitet 5 (111)	Att förlora tänder beskrivs som en traumatisk livshändelse som markerar en utförsbacke i livet. Man lever med ständig inre osäkerhet och är rädd för att tandlöshet eller dåliga tänder ska uppfattas som något komiskt. Detta kan beskrivas som en sorts oral överkänslighet. Man kan också anklaga sig själv för att man inte tagit ansvar för sin tandhälsa tidigare. Att vara tandlös upplevs som att vara mindervärdig som människa vilket också kan påverka könsrollen. Det förändrade utseendet känns som ett för tidigt åldrande, man sörjer sin förlorade ungdom och det kan vara så outhärdligt att man inte vill se sig i spegeln

Exemplet fortsätter på nästa sida

Exempel 8.2 Fortsättning.

Tabell 8.2.4 Exempel på tabellerade studier.

Author Year Reference Country	**Material method Analysis method**	**Informants**	**Results**
Fiske 1998 UK	Transcribed in-depth interview Qualitative approach	50 individuals (14 m, 36 fm) $\bar{x}$=69.9 years Toothless patients that seem well adapted to their dentures Dentures in 3 months–57 years $\bar{x}$=18.4 years	10 main themes: • bereavement • self-confidence • appearance • self-image • taboo • secrecy • prosthodontic privacy • behavioural change • premature ageing • lack of preparation
Trulsson 2002 Sweden	Transcribed in-depth interviews Grounded theory	18 individuals (8 m, 10 fm) 58–86 years $\bar{x}$=71 years Edentulous patients treated at Brånemark Clinic	3 categories with subcategories: Becoming an abnormal person: • lack of dental awareness earlier in life • feelings of shame and guilt • physical pain Loss of confidence: • physical suffering • feelings of shame • practical problems • decreased attractiveness Becoming the person I once was: • socially confident • feeling attractive again • good dental status • feelings of gratitude

	Summary	Study quality	Comments
	Loss of teeth like loss of any body part leads to a process of reactions: • to grieve • to cope with the acquired disability • to emotionally redefine the self	Moderate	The analysis is not fully described and could have been further developed This is an early qualitative study (1998) in this field and this may partly account for the methodological weaknesses
	Description of changes in self-image starting with the subjects' increasingly deteriorating dental status, followed by a period of having to live with and cope with a denture and, finally, living with a fixed prosthesis Motivation underlying the decision to undergo treatment with a fixed prosthesis seems to be a desire to restore not only oral function but also to regain earlier attractiveness, self-esteem and positive self-image	High	Relevant strategic selection of respondents The method is well described

Referenser

1. Sundhedsstyrelsen, Monitorering & Medicinsk Teknologivurdering. Kontrolforløb for gynækologiske kræftpatienter – en medicinsk teknologivurdering. København: Medicinsk Teknologivurdering; 2009.

2. Bower E, Scambler S. The contributions of qualitative research towards dental public health practice. Community Dent Oral Epidemiol 2007;35:161-9.

3. Noblit GW, Hare RD. Meta-ethnography: Synthesizing qualitative studies. Beverly Hills: Sage; 1988.

4. Centre for Reviews and Dissemination. Systematic Reviews: CRD's guidance for undertaking reviews in health care. University of York; 2008.

5. Kristensen FB, Sigmund H, editors. Health Technology Assessment Handbook. Copenhagen: Danish Centre for Health Technology Assessment, National Board of Health; 2007.

6. Larsson S. A pluralist view of generalization in qualitative research. International Journal of Research & Method in Education 2009;32:25-38.

7. Kuper A, Lingard L, Levinson W. Critically appraising qualitative research. BMJ 2008;337:a1035.

8. Malterud K. Kvalitativa metoder i medicinsk forskning: en introduktion. 2:a uppl. Lund: Studentlitteratur; 2009.

9. Polit DF, Beck CT. Essentials of nursing research. Appraising evidence for nursing practice. 7th ed. Philadelphia: Wolters Kluwer Health. Lippincott Williams & Wilkins; 2010.

10. Trulsson U, Engstrand P, Berggren U, Nannmark U, Bränemark P-I. Edentulousness and oral rehabilitation: experiences from the patient´s perspective. Eur J Oral Sci 2002;110:417-424.

11. Granskär M, Höglund-Nielsen B, red. Tillämpad kvalitativ forskning inom hälso- och sjukvård. Lund: Studentlitteratur; 2008.

12. Rapport F. Exploring the beliefs and experiences of potential egg share donors. J Adv Nurs 2003;Jul;43:28-42.

13. Holloway I, editor. Qualitative research in health care. Maidenhead: Open University Press; 2005.

14. Patton MQ. Qualitative research & evaluation methods. 3rd ed. London: SAGE; 2002.

15. Joanna Briggs Institute Reviewers' Manual: 2008 edition. Adelaide: The Joanna Briggs Institute; 2008.

16. Booth A. Formulating answerable questions. In: Booth A, Brice A, editors. Evidence based practice for information professionals: a handbook. London: Facet Publishing; 2004.

17. Evans D. Database searches for qualitative research. J Med Libr Assoc 2002;90:290-3.

18. Flemming K, Briggs M. Electronic searching to locate qualitative research: evaluation of three strategies. J Adv Nurs 2007;57:95-100.

19. Shaw RL, Booth A, Sutton A, Miller T, Smith JA, Young B, et al. Finding qualitative research: an evaluation of search strategies. BMC Med Res Methodol 2004;4:5.

20. Booth A. Chapter 3: Searching for studies. In: Noyes J, Booth A, Hannes K, Harden A, Harris J, Lewin S, et al, editors. Supplementary guidance for inclusion of qualitative research in Cochrane systematic reviews of interventions. Version 1 (updated August 2011): ed: Cochrane Collaboration Qualitative Methods Group; 2011.

21. SBU. Metoder för behandling av långvarig smärta. En systematisk litteraturöversikt. Stockholm: Statens beredning för medicinsk utvärdering (SBU); 2006. SBU-rapport nr 177/2. ISBN 91-85413-09-7.

22. Willman A, Stoltz P, Bahtsevani C. Evidensbaserad omvårdnad: en bro mellan forskning och klinisk verksamhet. 3:dje uppl. Lund: Studentlitteratur; 2011.

23. Pope C, Mays N, editors. Qualitative research in health care. 3rd ed. London: BMJ Books; 2006.

24. Lincoln YS, Guba EG. Naturalistic inquiry. Beverly Hills: Sage; 1985.

25. SBU. Schizofreni. Läkemedelsbehandling, patientens delaktighet och vårdens organisation. En systematisk litteraturöversikt. Stockholm: Statens beredning för medicinsk utvärdering (SBU); 2012. SBU-rapport nr 213. ISBN 978-91-85413-50-8.

26. Howell Major C, Savin-Baden M. An introduction to qualitative research synthesis. London: Routledge publishing; 2010. ISBN 10: 0-415-56286-4.

27. Timulak L. Meta-analysis of qualitative studies: a tool for reviewing qualitative research findings in psychotherapy. Psychother Res 2009;19:591-600.

28. SBU. Tandförluster. En systematisk litteraturöversikt. Stockholm: Statens beredning för medicinsk utvärdering (SBU); 2010. SBU-rapport nr 204. ISBN 978-91-85413-40-9.

29. Corbin J, Strauss A. Basics of qualitative research: techniques and procedures for developing grounded theory. 3rd ed. Thousand Oaks: SAGE; 2008.

30. Winther Jørgensen M, Phillips L. Diskursanalys som teori och metod. Lund: Studentlitteratur; 2000.

9. Sammanvägning av resultat

VERSION 2011:1.1

Vid utvärderingar inom hälso- och sjukvård gäller det att bedöma vilken alternativ intervention som har störst effekt för ett givet problematiskt tillstånd. Om det finns flera utvärderingar av de alternativa interventionerna, så behöver studieresultaten vägas samman på något sätt. De sammanvägda resultaten kan ingå i en evidensprofil (se Kapitel 10 om GRADE) och därefter fungera som en del i ett beslutsunderlag inom evidensbaserad medicin [1]. Om sammanvägningen görs med hjälp av statistiska metoder kallas den för metaanalys; om statistiska metoder inte används brukar man tala om narrativa sammanvägningar. Metaanalyser används oftast avseende randomiserade studier (RCT) av alternativa interventioner. De förekommer dock även vid sammanvägningar av andra typer av studier, t ex inom diagnostik och psykometri. Syftet med detta kapitel är att ge en orientering om följande:

- vad det innebär att göra en metaanalys
- problem som är aktuella vid metaanalys samt strategier för att hantera problemen:
  - publikationsbias: ”funnel plots” och ”trim and fill”
  - bristande samstämmighet (heterogenitet): subgruppsanalys, ”random effects model” respektive ingen sammanvägning
- metaanalys vid observationsstudier, diagnostik och psykometri.

Alla resultat har inte samma tyngd

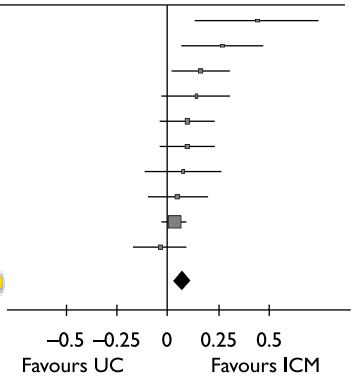
Metaanalys innebär normalt att man räknar fram ett slags genomsnitt avseende flera studieresultat för att skatta en enda ”sann” effekt. Alla enskilda resultat har dock normalt inte samma tyngd i sammanvägningen. Intuitivt kan man tycka att små studier borde väga mindre än stora studier vid sammanvägningen. Detta stämmer också i viss mån. Den relativa tyngd som varje resultat har beror normalt sett på antalet individer i studien; ju fler individer, desto tyngre blir resultatet i sammanvägningen. Om man ska vara korrekt är det stickprovsfördelningens spridning (standardfelet) som avgör, ju mindre spridning, desto större tyngd (denna spridning minskar om antalet individer ökar) [2].

Ett vanligt sätt att gestalta metaanalysen är en så kallad ”forest plot” (skogsdiagram). Denna innehåller bl a skattade effektstorlekar för varje studie, en sammanvägd effektstorlek samt konfidensintervall för såväl de enskilda effekterna som för den sammanvägda effekten. I Figur 9.1 visas en ”forest plot”, med resultaten av en intervention för hemlösa personer med psykisk funktionsnedsättning och mer eller mindre grava missbruksproblem [3–12]. Interventionen består av ett program kallat intensiv ”case management” (ICM) medan kontrollalternativet består av standardvård (UC för ”usual care”).

Effektmåttet är riskskillnad (”risk difference”)[1]. Riskskillnad anger här hur många procent fler i interventionsgruppen som klarat av eget boende vid 12-månadersuppföljningen jämfört med kontrollgruppen, alltså skillnaden mellan två proportioner. Man brukar använda ordet ”risk” även om det rör sig om positiva händelser som t ex tillfrisknande. Resultatet från varje enskild studie benämns enligt försteförfattaren, de horisontella linjerna visar konfidensintervallen och rektangeln i mitten visar vilken effektstorleken är.

Study or subgroup	Intensive case management Events	Total	Usual care Events	Total	Weight	Risk Difference IV, Fixed, 95% CI
Lipton 1988	16	20	5	14	1.8%	0.44 [0.14, 0.75]
Bond 1990	30	45	17	43	4.1%	0.27 [0.07, 0.47]
Morse 1992	33	37	48	66	7.7%	0.16 [0.02, 0.31]
Lehman 1997	37	64	27	62	5.6%	0.14 [–0.03, 0.32]
Cox 1998	43	108	25	83	9.1%	0.10 [–0.04, 0.23]
Sosin 1995	49	70	73	121	8.7%	0.10 [–0.04, 0.23]
Morse 2006	25	54	19	49	4.6%	0.08 [–0.12, 0.27]
Rosenheck 2003	41	62	69	113	7.6%	0.05 [–0.10, 0.20]
Conrad 1998	162	178	158	180	41.2%	0.03 [–0.03, 0.10]
Clarke 2000	89	114	40	49	9.5%	–0.04 [–0.17, 0.10]
Total (95% CI)		**752**		**780**	**100.0%**	**0.07 [0.03, 0.12]**
Total events	525		481			

Heterogeneity: Chi²=15.94, df=9 (p=0.07); I²=44%
Test for overall effect: Z=3.57 (p=0.0004)



CI = Confidence interval; ICM = Intensive case management; UC = Usual care

Figur 9.1 Exempel på metaanalys (”forest plot”) – intensiv ”case management” (ICM) mot standardvård (UC).

För Lehman och medarbetares studie är resultatet följande: riskskillnaden är 14 procent, alltså 14 procent fler i interventionsgruppen än kontrollgruppen hade ett eget stabilt boende vid 12-månadersuppföljningen. Konfidensintervallet, från –3 till 32 procent, överlappar emellertid 0-linjen. Detta innebär att skillnaden ligger inom den statistiska felmarginalen. Resultatet är med andra ord inte statistiskt signifikant. Diamanten (romboiden) längst ner visar den sammanvägda effekten samt konfidensintervallet för den sammanvägda effekten: en riskskillnad på 7 procent och ett konfidensintervall från 3 till 12 procent.

I kolumnen med rubriken ”weight” framgår vilken tyngd respektive resultat har i sammanvägningen. Det lättaste resultatet (knappt 1,8 procent) kommer från en studie av Lipton och medarbetare medan det resultat som väger tyngst har presenterats i en studie av Conrad och medarbetare (41,2 procent). Notera att ett resultat väger tyngre ju kortare konfidensintervallet är. Detta beror på att ju större standardfelet är, desto längre blir konfidensintervallet.

[1] Det är vanligt att man istället använder oddskvot eller riskkvot vid medicinska utvärderingar beroende på de statistiska egenskaper dess mått har. Vi har valt riskskillnad eftersom detta mått är enklast att förstå.

Figur 9.1 kan illustrera varför man gör metaanalyser. För det första resulterar metaanalysen i *en* sammanvägd effekt från de tio ingående resultaten (diamanten längst ner i Figur 9.1). Det underlättar tolkningen av resultaten vid en utvärdering om man har *en* effekt med *ett* konfidensintervall istället för tio olika effekter med tio olika konfidensintervall. För det andra ökar precisionen i skattningen av effekten normalt sett jämfört med precisionen i de enskilda resultaten. Det betyder att risken att man missar en ”sann” effekt pga att antalet ingående individer är för litet minskar[2].

Det finns emellertid några problem som gör att den sammanvägda effekten i Figur 9.1 inte alltid är en tillförlitlig skattning av den ”sanna” effekten. För det första kan det vara så att de resultat som ingår i metaanalysen inte utgör ett representativt urval pga av ett problem som kallas publikationsbias. Vanligtvis innebär detta att den skattade effekten är något för stor. För det andra kan resultaten baseras på studier där åtminstone några studier inte är tillräckligt lika de andra avseende t ex populationens sammansättning, lokal kontext (sammanhang), interventionernas exakta innehåll, kontrollvillkoren, sättet att mäta effekterna, samt studiedesign. Detta problem brukar kallas klinisk heterogenitet [13] och kan ta sig uttryck i såväl en över- som en underskattning av den ”sanna” effekten. I följande avsnitt kommer vi att visa hur metaanalys kan användas för att hantera sådana problem, först publikationsbias och därefter heterogenitet.

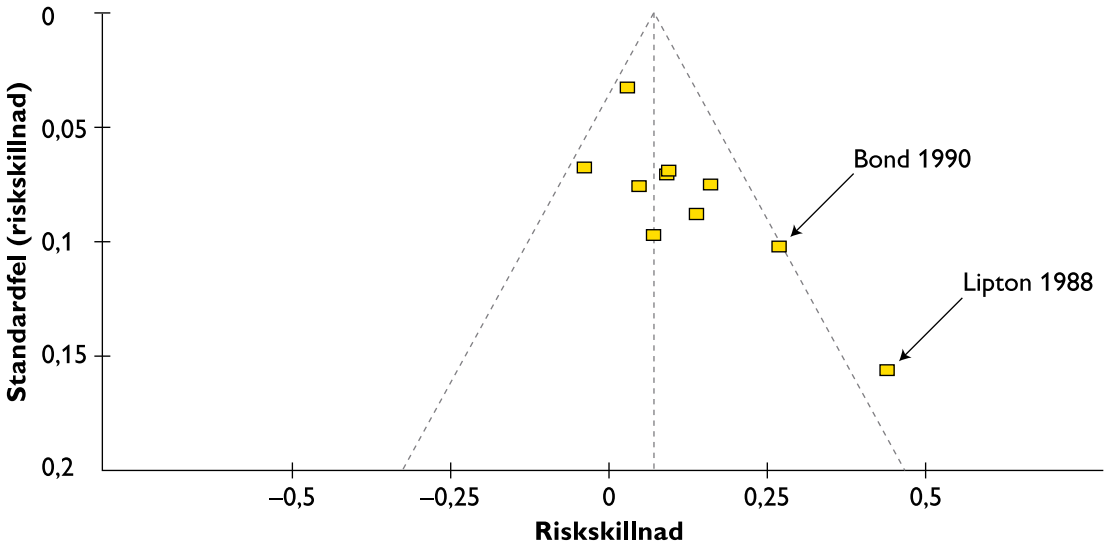
Publikationsbias och ”funnel plots”

I Figur 9.2 har resultaten från Figur 9.1 omgestaltats till en ”funnel plot” (trattdiagram). Effektstorleken visas på den horisontella axeln medan spridningen (standardfelet) visas på den vertikala axeln. Observera att den vertikala axelns värden är omvända så att ju högre upp på axeln ett resultat finns, desto lägre är spridningen. Kvadraten längst ner till höger visar resultatet från Lipton och medarbetare med en effekt på 44 procent och med den största spridningen av alla ingående studier. Den streckade triangeln är en hjälp för att visuellt kunna tolka resultatet. Den lodräta mittenlinjen visar var den sammanvägda effekten på 7 procent ligger.

Modellen bygger bl a på två antaganden: (a) att resultat från stora studier (med liten spridning) är lättare att publicera än resultat från små och (b) att resultat med en stor effekt till förmån för den utvärderade interventionen är lättare att publicera än resultat som inte är signifikanta eller som talar emot interventionen [2,14]. Publiceringssvårigheten kan ta sig uttryck i att de små icke-positiva resultaten aldrig blir publicerade, att publiceringen tar längre tid eller att publicering sker i tidskrifter som inte indexeras i referensdatabaser (och kan därmed vara svåra att hitta). Medvetenheten om dessa

[2] Risken för typ-2 fel eller β-fel minskar vid metaanalys eftersom den statistiska teststyrkan ökar.

publiceringsproblem kan även leda till en selektiv rapportering inom varje enskild studie på så sätt att man endast rapporterar de statistiskt signifikanta resultat som talar för interventionen och undviker att rapportera övriga resultat. Ibland talar man om rapporteringsbias, något som inte riktigt är samma sak som publikationsbias. Rapporteringsbias innebär en tendentiös rapportering inom en och samma studie, alltså en benägenhet att endast rapportera de resultat som stödjer interventionen. Om rapporteringsbias är mer vanligt inom småstudier än inom stora, tar sig detta samma uttryck som publiceringsbias. Det bör även nämnas att det kan finnas ekonomiska intressen bakom denna typ av selektiv rapportering om de som utvärderar interventionen kan ha nytta av att den framstår som effektiv.



Figur 9.2 Trattdiagram (”funnel plot”) – tecken på publikationsbias.

Om ovanstående antaganden stämmer, så borde det finnas relativt få studieresultat i den vänstra nedre hörnan av triangeln (alltså små studier som talar emot interventionen eller som inte är statistiskt signifikanta). Om det inte fanns något publikationsbias, borde resultaten fördela sig symmetriskt kring den sammanvägda och skattade effekten. I Figur 9.2 finns tecken på publikationsbias. Detta betyder att riskskillnaden på 7 procent kan vara en överskattning av den ”sanna” effekten.

För att få en bild av hur mycket effekten överskattas kan man plocka bort de mest extrema resultaten (”trim”) till förmån för interventionen och därefter räkna fram en ny effektstorlek. För att det sammanvägda konfidensintervallets längd inte ska överskattas så kan nya hypotetiska resultat läggas till (”fill”). Detta sätt att hantera publikationsbias har utvecklats till en statistisk metod kallad ”trim and fill” där man med hjälp av en iterativ process räknar fram ett resultat där publikationbias har hanterats [2]. Om t ex resultatet från Lipton och medarbetares studie tas bort, så förändras inte den skattade effekten på

7 procent, men konfidensintervallets övre gräns minskar från 0,12 till 0,11. Om även Bond och medarbetares studie tas bort, minskar effekten till 6 procent och konfidensintervallet går från 0,02 till 0,10; resultatet är alltjämt statistiskt signifikant. Ovanstående metodologiska övningar kan ge en bild av resultatens konsistens och hur stort ett publikationsbias skulle kunna vara i detta exempel.

Bristande samstämmighet kan tydliggöras och undersökas

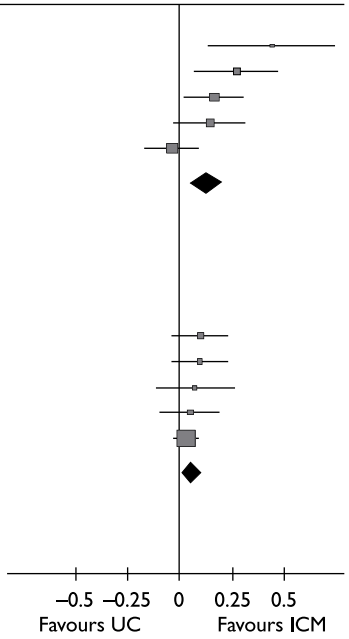
I detta avsnitt kommer vi att visa hur problemet med heterogenitet kan hanteras med hjälp av metaanalys [2,13]. Även om alla resultat utom ett i Figur 9.1 uppvisar en positiv effekt så är inte resultaten samstämmiga. Exempelvis så varierar effektstorleken en hel del, från 44 procent (Lipton och medarbetare) till minus 4 procent (Clarke och medarbetare). Det går att kvantifiera denna bristande samstämmighet med olika mått på heterogenitet såsom I^2 och Q. Q är ett vägt mått som baseras på de avvikelser som varje enskilt resultat har från den sammanvägda effekten. Med hjälp av ett χ^2-test framgår att heterogeniteten är statistiskt signifikant i exemplet eftersom p=0,07 <0,10 (som tumregel brukar 0,10 användas som gräns av försiktighetsskäl). Hur stor andel av den totala variansen som förklaras av variansen mellan de enskilda resultaten fångas upp av I^2, 44 procent i fallet ovan. Annorlunda uttryckt, I^2 utgör andelen av den totala variansen som förklaras av att det finns reella skillnader i effektstorlekar studierna emellan. Enligt en tumregel brukar I^2 benämnas på följande sätt: låg heterogenitet = 0,25, måttlig heterogenitet = 0,50 och hög heterogenitet = 0,75 [2].

Anta att de olika resultaten bygger på studier som är mycket lika varandra avseende interventioner, kontrollvillkor, utvärderingsdesign och effektmått. Anta vidare att populationerna varierar från ett resultat till ett annat, men att positiva effekter trots detta uppvisar stor samstämmighet. Under sådana omständigheter tyder resultatet sammantaget på att interventionens skattade effektivitet är förhållandevis stabil oavsett subgrupper inom populationen (allt annat lika). I Figur 9.1 är resultaten emellertid inte samstämmiga vilket visar sig i den statistiska heterogeniteten.

Det kan emellertid finnas kliniska och metodologiska förklaringar till den bristande samstämmigheten. En möjlighet är att skilda patientgrupper reagerar olika på interventionen ICM. ICM har i första hand utvecklats för personer med psykisk funktionsnedsättning, t ex schizofreni (MI för ”mental illness”). Det kan därför vara så att ICM fungerar annorlunda för patienter vars huvudsakliga problem är tungt drogmissbruk (SA för ”substance abuse”) eller både tungt drogmissbruk och psykisk funktionsnedsättning (MISA). En strategi att hantera heterogeniteten skulle därför kunna vara att analysera betydelsen av olika subgrupper.

I Figur 9.3 har resultaten delats upp i två subgrupper men någon total sammanvägning har inte gjorts. Med denna gruppindelning framgår att det inte finns någon heterogenitet inom SA/MISA-gruppen medan den t o m ökar inom MI-gruppen. Detta skulle kunna tyda på att ICM fungerar olika i de två grupperna av patienter, sämre i SA/MISA och bättre i MI-gruppen jämfört med UC. Andelen av den totala variansen som förklaras av de två subgrupperna är mer än måttligt stor (60,7 procent), varför uppdelning i subgrupper kan vara lämplig. Eftersom heterogeniteten i MI-gruppen ökar och skillnaden mellan subgrupperna inte är statistiskt signifikant (p=0,11), kanske det är lämpligt att gå vidare med ytterligare subgrupper inom MI-gruppen eller att redovisa resultaten separat för de enskilda studierna. Det kan dock finnas andra alternativ för att förklara den bristande samstämmigheten.

Study or subgroup	Intensive case management Events	Total	Usual care Events	Total	Weight	Risk Difference IV, Fixed, 95% CI
MI						
Lipton 1988	16	20	5	14	6.2%	0.44 [0.14, 0.75]
Bond 1990	30	45	17	43	14.4%	0.27 [0.07, 0.47]
Morse 1992	33	37	48	66	26.9%	0.16 [0.02, 0.31]
Lehman 1997	37	64	27	62	19.4%	0.14 [–0.03, 0.32]
Clarke 2000	89	114	40	49	33.1%	–0.04 [–0.17, 0.10]
Subtotal (95% CI)		**280**		**234**	**100.0%**	**0.13 [0.05, 0.20]**
Total events	205		137			
Heterogeneity: Chi²=12.15, df=4 (p=0.02); I²=67%						
Test for overall effect: Z=3.26 (p=0.001)						
SA/MISA						
Cox 1998	43	108	25	83	12.8%	0.10 [–0.04, 0.23]
Sosin 1995	49	70	73	121	12.3%	0.10 [–0.04, 0.23]
Morse 2006	25	54	19	49	6.5%	0.08 [–0.12, 0.27]
Rosenheck 2003	41	62	69	113	10.7%	0.05 [–0.10, 0.20]
Conrad 1998	162	178	158	180	57.8%	0.03 [–0.03, 0.10]
Subtotal (95% CI)		**472**		**546**	**100.0%**	**0.05 [0.00, 0.10]**
Total events	320		344			
Heterogeneity: Chi²=1.25, df=4 (p=0.87); I²=0%						
Test for overall effect: Z=2.16 (p=0.03)						
Test for subgroup differences: Chi²=2.54, df=1 (p=0.11); I²=60.7%						



CI = Confidence interval; ICM = Intensive case management; MI = Mental illness; SA/MISA = Substance abuse/Mental illness and substance abuse; UC = Usual care

Figur 9.3 Subgrupper – psykisk funktionsnedsättning och drogmissbruk, intensiv ”case management” (ICM) mot standardvård (UC).

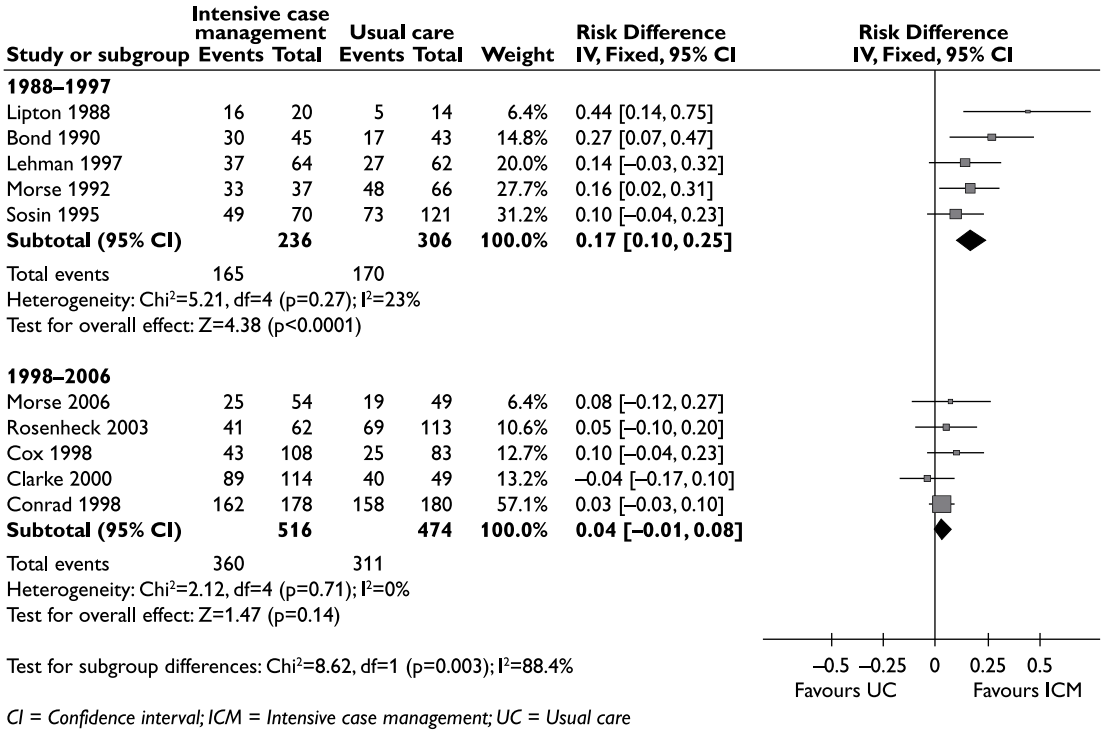
Bakom heterogeniteten kan det finnas ett metodologiskt problem. Detta problem kan uppträda när kontrollvillkoret utgörs av standardvård och den utvärderade interventionen består av en sammansättning av flera mer eller mindre verksamma komponenter. Detta problem har att göra med att komponenter, som ingår i den nya och kanske mer effektiva interventionen, börjar spridas och integreras som delar i interventioner som ingår i standardvården (en slags kontaminering). Om detta stämmer borde effekten av ICM i jämförelse med UC bli allt mindre över tid eftersom UC blir allt mer lik ICM

över tid. I Figur 9.4 har nya subgrupper bildats där resultaten delats upp i två hälfter i enlighet med medianen (mellan år 1997 och 1998) för det tidsspann som omfattas. Med denna nya indelning försvinner heterogeniteten i båda subgrupperna, skillnaden mellan subgrupperna blir statistiskt signifikant (p=0,003) och andelen av den totala variansen som förklaras av de två subgrupperna blir hela 88,4 procent.

Om antagandet om kontaminering stämmer, borde de minskande effekterna över tid i första hand bero på att UC klarar sig allt bättre samtidigt som ICM-gruppens resultat ligger på ungefär samma nivå över tid. Om man summerar samtliga individer i respektive grupp från de två tidsintervallen blir resultatet följande:

- Av de deltagare som fått UC befann sig 56 procent (170/306=0,56) i stabilt boende vid 12-månadersuppföljningen under åren 1988–1997. För perioden efter 1998–2006 var motsvarande andel för UC-gruppen 66 procent (311/474=0,66). Detta innebär en förbättring på 10 procentenheter.

- Av de deltagare som fått ICM återfanns 70 procent i stabilt boende vid 12-månadersuppföljningen för båda tidsintervallen (165/236=0,70 och 360/516=0,70).



CI = Confidence interval; ICM = Intensive case management; UC = Usual care

Figur 9.4 Subgrupper – studier år 1988–1997 samt 1998–2006, intensiv ”case management” (ICM) mot standardvård (UC).

Dessa två resultat pekar på att kontrollgruppen kan ha kontaminerats över tid. I detta fall när det finns en kontinuerlig variabel som skulle kunna förklara effektstorlekens variation så skulle man även kunna använda sig av metaregression som analysverktyg istället för två tidsperioder [2].

Figur 9.3 och 9.4 exemplifierar vad subgruppsanalys kan innebära som strategi att hantera problemet med heterogenitet, dvs med bristande samstämmighet. Det troliga är att den bristande samstämmigheten har flera orsaker och övningarna ovan visar att såväl en heterogen patientpopulation som metodologiska problem kan ligga bakom. Det kan dock finnas ytterligare orsaker.

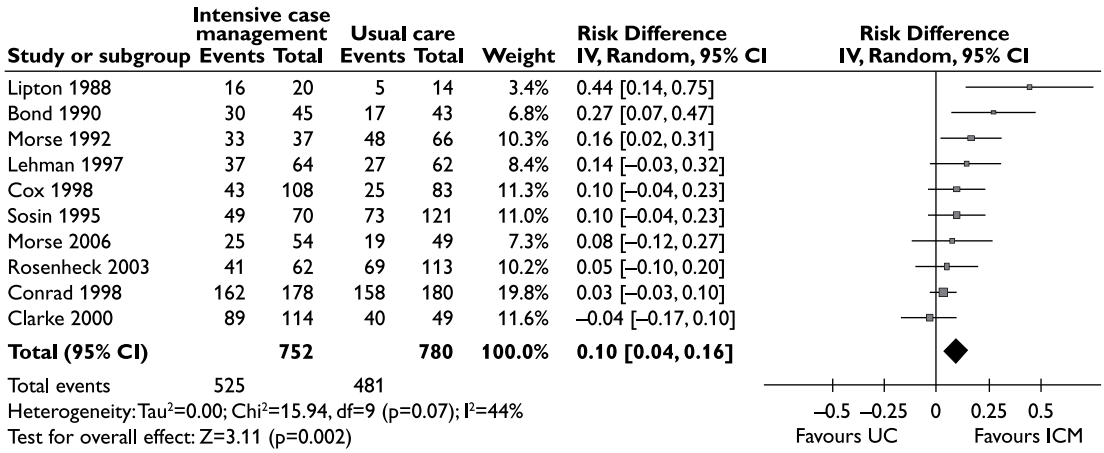
Bristande samstämmighet kan inkluderas i metaanalysmodellen

Vi har hittills använt en ”fixed effect model” (FEM) i Figur 9.1–9.4 [2]: under ”Risk Difference” står det ”Fixed”. Denna modell bygger på antagandet att samtliga resultat utgör slumpmässiga urval från en och samma population där det finns en enda ”sann” effekt. Ett annat ganska vanligt sätt att hantera heterogenitet är emellertid att använda sig av en annan modell som bygger på andra antaganden. Denna alternativa modell kallas ”random effects model” (REM) [2]. När denna modell används utgår man från att varje studieresultat baseras på slumpmässiga urval från flera populationer av resultat med en egen ”sann” effekt för varje studie. I praktiken betyder detta att små avvikande studier kommer att väga mer med ”random” än med ”fixed effects model”. Det kan nämnas att ju mindre heterogena resultat, desto mindre blir skillnaderna i resultat mellan modellerna.

I Figur 9.5 visas hur resultaten förändras jämfört med Figur 9.1 då REM används. För det första ökar effekten från 7 till 10 procent samt att konfidensintervallet både förskjuts och blir längre: 0,04 till 0,16 istället för 0,03 till 0,12. Vidare bör det noteras att det tyngsta resultatet i Figur 9.1 – från studien av Conrad och medarbetare – minskar från 41,2 till 19,8 procent samt att det lättaste resultatet i Lipton och medarbetares studie ökar från 1,9 till 3,4 procent.

Att dela upp resultaten i subgrupper (Figur 9.3 och 9.4) eller inkludera heterogeniteten i metaanalysmodellen (Figur 9.5) är olika sätt att hantera bristande samstämmighet. Vad dessa alternativa strategier innebär blir tydligt när man tolkar resultaten. Resultaten i Figur 9.5, en effekt på 10 procent i riskskillnad, bygger på antagandena att det finns tio olika populationer med vardera en egen sann effekt för varje studie. De tio skilda resultaten antas utgöra slumpmässiga urval av studier från dessa respektive populationer. Den sammanvägda effekten på 10 procent är därför inte en skattning av *en* sann effekt utan

en skattning av medelvärdet i en fördelning av skattade ”sanna” effekter. Uppdelningen i subgrupper (Figur 9.3 och 9.4) istället för att använda REM innebär antaganden om att det finns två populationer, en för vardera subgruppen, och två ”sanna” effekter. Dessa populationer bedöms vara för olika för att det ska vara meningsfullt att inkludera resultat från dem i samma sammanvägning.



CI = Confidence interval; ICM = Intensive case management; UC = Usual care

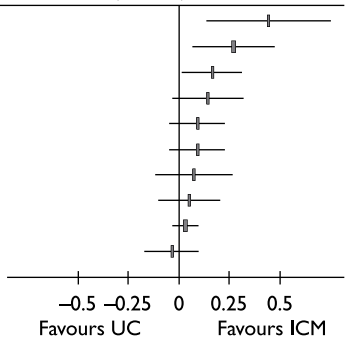
Figur 9.5 ”Random effects model” – intensiv ”case management” (ICM) mot standardvård (UC).

Stor klinisk heterogenitet och ingen statistisk sammanvägning

Varje enskilt resultat baseras på studier som kan vara olika varandra avseende patientpopulationer (t ex sammansättning, riskfaktorer), interventioner (t ex innehåll inklusive tilläggsbehandlingar, implementering), kontrollvillkor (t ex innehåll inklusive tilläggsbehandlingar, implementering), effektmått (t ex definitioner, mätmetoder, uppföljningstid) samt studiedesign (t ex allokeringsmetoder, hantering av behandlingsavbrott). Om olikheterna är för stora, kan man helt enkelt avstå från att väga samman resultaten till en enda skattning av effektstorleken. Att sammanfatta resultaten i en ”forest plot” kan emellertid ändå vara till hjälp när resultaten sedan ska tolkas (Figur 9.6).

Samtliga resultat med samma statistiska effektmått (riskskillnad) inklusive konfidensintervall finns med i figuren. Detta gör materialet överskådligt jämfört med om effekterna skulle redovisas i separata figurer eller enbart i texten. Genom att inte räkna fram en sammanvägd effekt kan man markera att detta inte skulle vara lämpligt. Om materialet är alltför komplext och heterogent, skulle en sammanvägning kunna ge en vilseledande tilltro till en precision som inte är möjlig. Sammanvägningar av resultaten, såsom de presenteras i Figur 9.6, kan därför inte vara statistiska utan istället narrativa. Detta betyder att man måste tolka och sammanfatta hela bilden som framträder i Figur 9.6 med ord.

Study or subgroup	Intensive case management Events	Total	Usual care Events	Total	Risk Difference IV, Fixed, 95% CI
Lipton 1988	16	20	5	14	0.44 [0.14, 0.75]
Bond 1990	30	45	17	43	0.27 [0.07, 0.47]
Morse 1992	33	37	48	66	0.16 [0.02, 0.31]
Lehman 1997	37	64	27	62	0.14 [–0.03, 0.32]
Cox 1998	43	108	25	83	0.10 [–0.04, 0.23]
Sosin 1995	49	70	73	121	0.10 [–0.04, 0.23]
Morse 2006	25	54	19	49	0.08 [–0.12, 0.27]
Rosenheck 2003	41	62	69	113	0.05 [–0.10, 0.20]
Conrad 1998	162	178	158	180	0.03 [–0.03, 0.10]
Clarke 2000	89	114	40	49	–0.04 [–0.17, 0.10]



CI = Confidence interval; ICM = Intensive case management; UC = Usual care

Figur 9.6 ”Forest plot” utan sammanvägning.

Analysverktyg eller del i evidensprofil

Metaanalys kan användas på olika sätt. I exemplen ovan har metaanalysen fungerat som ett analysverktyg med vars hjälp man får en bättre förståelse för de data man arbetar med. När väl slutrapporten skrivs och underlaget ska tabuleras i GRADE och bilda en evidensprofil (Kapitel 10), bör man välja de metaanalyser som ska ingå med omsorg. Detta gäller vilka studieresultat som ska ingå, val av modell (”fixed effects model” eller ”random effects model”), eventuella subgrupper samt om man överhuvudtaget ska göra någon sammanvägning. Även val av statistiska effektmått (oddskvot, riskkvot, riskskillnad, hasardkvot m m) behöver motiveras. Dessa mått har olika statistiska egenskaper och är inte alltid lika lämpliga. I exemplen ovan har riskskillnad valts av pedagogiska skäl eftersom riskskillnad är lätt att förstå intuitivt.

Dessa val kan spela stor roll då beslut ska fattas i valet mellan alternativa interventioner inom hälso- och sjukvård. Om Figur 9.1 eller 9.5 väljs, talar resultaten för ICM framför UC (allt annat är lika) oavsett vilken subgrupp det rör sig om (psykisk funktionsnedsättning eller tungt drogmissbruk med eller utan psykisk funktionsnedsättning). Om hänsyn tas till ett eventuellt publikationsbias (Figur 9.2), förändrar inte detta bilden, även om de förväntade effekterna blir något mindre. Ett val av Figur 9.3 skulle innebära att ICM är att föredra om psykiskt funktionshinder utgör huvudproblemet. Om tungt drogmissbruk finns med i bilden är inte detta val lika klart (allt annat lika). Om man väljer metaanalysen i Figur 9.4 som del av evidensprofilen, är det tveksamt om ICM är att föredra framför UC, oavsett om psykiskt funktionshinder eller tungt drogmissbruk utgör huvudproblemet. Det verkar ju som om UC har förbättrats så pass mycket under det senaste decenniet att det inte längre verkar finnas någon skillnad jämfört med ICM.

Anta att Figur 9.6 används i det slutgiltiga underlaget som en del i en evidensprofil. I detta fall har resultaten bedömts komma från studier vilka är alltför olika för att en sammanvägning ska vara meningsfull. Det centrala i detta fall är vilket eller vilka

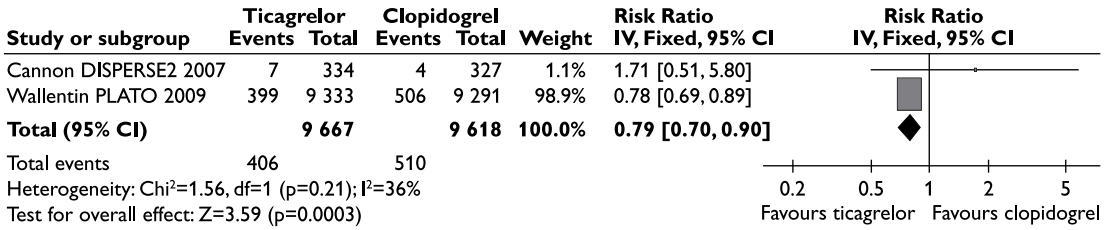
resultat som är mest relevanta för beslutsfattande inom den svenska praktiken (jämför överförbarhet i GRADE). Alla tio resultat i Figur 9.6 kanske inte är lika relevanta om patientpopulation, intervention, kontrollalternativ, effektmått och design beaktas i detalj. Kanske Morse och medarbetare är den studie som bäst fångar de alternativ som praktiken står inför, kanske något eller några andra studieresultat är mest relevanta. Vilket eller vilka resultat som då väljs avgör förstås vilket beslut som stöds på samma sätt varför detta val måste motiveras på ett systematiskt och transparent sätt.

Det bör betonas att de forskningsfrågor man försöker besvara med hjälp av metaanalys kan vara mycket olika. Detta beror dels på hur frågan specificerats samt på hur det aktuella forskningsfältet ser ut. Ovanstående exempel kommer från ett forskningsfält präglat av många studier med få deltagare, komplexa och ofta otillräckligt beskrivna interventioner och kontrollvillkor samt inte alltid tillförlitliga effektmått. Forskningsfrågan är även förhållandevis vid. Med en mer snävt avgränsad fråga inom ett metodologiskt starkare forskningsfält kan det se ut på ett helt annat sätt.

Anta att man vill veta hur totalmortaliteten påverkas för personer med akuta koronara syndrom av två alternativa trombocythämmande läkemedel: ticagrelor och clopidogrel. För denna fråga finns i skrivande stund endast två randomiserade studier att tillgå [15,16], men det är två välgjorda och stora studier. Den ena, PLATO, omfattar drygt 18 000 deltagare från över 740 olika center fördelade över nästan hela världen, medan den andra, DISPERSE2, hade 990 deltagare från 132 center. Resultaten visas i Figur 9.7 i form av relativa risker (andelen döda i ticagrelorgruppen dividerat med andelen döda i clopidogrelgruppen)[3].

Resultaten i den större studien är statistiskt signifikanta och talar för ticagrelor, medan det i den mindre studien finns en icke-signifikant och minimal överrisk (tre personer) för ticagrelorgruppen. Även om det inte föreligger någon statistisk heterogenitet, kan man ändå konstatera att de två studierna ger olika budskap. Det finns dock några kliniska och metodologiska skillnader mellan de två studierna. För det första är den totala andelen döda i PLATO 4,9 procent mot 1,7 procent i DISPERSE2, vilket tyder på att patienterna kan ha varit något friskare i DISPERSE2. I DISPERSE2 var målgruppen patienter med akuta koronara syndrom utan ST-segment-elevation medan målgruppen för PLATO även inkluderade denna grupp. För det andra var uppföljningstiden 12 månader i PLATO medan den endast var 3 månader i DISPERSE2. Antalet händelser (”events”) är också betydligt färre i DISPERSE2-studien.

[3] I ticagrelorstudien redovisas effekten som hasardkvot vilket är ett bättre alternativ än riskkvot eftersom tiden till händelsen beaktas. Eftersom dessa uppgifter inte finns i den mindre har vi använt riskkvot.



CI = Confidence interval

Figur 9.7 Ticagrelor mot clopidogrel.

Sammantaget kan detta betyda att studierna bedöms vara för olika för att vägas samman som underlag i en evidensprofil. Om man bedömer att en uppföljningstid på 12 månader eller längre krävs för tillförlitliga resultat, kan man välja bort DISPERSE2 från evidensprofilen. Det är slutligen så att rent statistiskt spelar DISPERSE2 ingen roll, eftersom den skattade effekten och konfidensintervallet inte påverkas märkbart om DISPERSE2 tas med eller tas bort. Studien väger endast 1 procent.

Det kan verka poänglöst att genomföra en metaanalys med endast två studier som kanske är för olika för att vägas samman. Som analysverktyg kan metaanalysen ändå ha sin roll. Skillnaden mellan de två studieresultaten blir tydliga. Detta kan göra att man blir uppmärksam på kliniska och metodologiska skillnader man kanske inte uppmärksammat tidigare. Slutligen blir studieresultatens relativa statistiska vikt tydlig. Allt detta kan vara till hjälp när man slutligen bedömer vad som ska ingå i evidensprofilen. Om de två studierna bedöms vara tillräckligt lika bör en sammanvägning ingå i evidensprofilen för att förenkla presentation och tolkning (se Kapitel 10 om GRADE).

Metaanalyser från observationsstudier

Det går att göra metaanalyser avseende resultat från *observationsstudier*, även om det är mindre vanligt än för randomiserade studier och ofta mer arbetskrävande. Grundprincipen är emellertid samma. Man väger samman effekter där interventioner jämförs med kontrollvillkor. Det finns dock ett antal praktiska och principiella problem som gör det hela svårare och mer arbetskrävande än då randomiserade studier används. Observationsstudier präglas av stor variation avseende metodologiska upplägg. Variationen kan t ex bero på om det finns en matchad jämförelsegrupp (kontrollgrupp) vid baslinjen (mätningar före intervention) eller om man skapar en matchning i efterskott genom någon form av multivariat metodik, antalet jämförelsegrupper och vid hur många tidpunkter mätningar görs. Campbell och Stanley [17] lyfter fram 14 varianter medan Shadish et al [18] beskriver ett 20-tal studieupplägg vilka delats in i fyra olika kategorier: (a) observationsstudier som såväl saknar jämförelsegrupp som mätningar vid baslinjen, (b) observationsstudier som har såväl jämförelsegrupp som mätningar vid baslinjen, (c) avbrutna tidsserier samt (d) avbruten regressionsdesign (”regression discontinuity design”).

Samtliga alternativa studieupplägg bör inkludera någon modell med vars hjälp man försöker hantera problem med risk för selektionsbias. Selektionsbias kan uppkomma då interventions- och kontrollgrupper inte är tillräckligt lika avseende t ex risk och skyddsfaktorer. För att metaanalyser baserade på observationsstudier ska vara praktiskt möjliga, krävs att data finns tillgängliga i ett format där interventionsgruppen ställs mot en jämförelsegrupp efter justeringar för eventuella skillnader. Man kan matcha kontrollgruppen mot interventionsgruppen vid baslinjen med stöd av t ex kända risk- och skyddsfaktorer med syftet att grupperna ska vara så lika som möjligt. I andra fall försöker man skapa likvärdighet i efterhand med hjälp av statistiska modeller.

Om syftet med studien är att utvärdera en intervention i jämförelse med ett kontrollalternativ kan det vara möjligt att använda studieresultaten i en metaanalys. Om huvudsyftet inte varit en sådan utvärdering utan istället att testa en kausal modell kan det vara svårare att använda resultaten i en metaanalys, speciellt om det inte finns tillräckligt med statistisk information (t ex antal individer, medelvärden, spridningsmått).

Att använda metaanalysen baserad på observationsstudier som ett analysredskap behöver inte vara förenat med några större principiella problem; det kan t ex handla om att få grepp om heterogeniteten. Att använda de statistiska sammanvägningarna som en del i en evidensprofil kan emellertid vara mer riskabelt avseende resultat från observationsstudier än från randomiserade studier. De justeringar man gjort kan avse olika bakgrundsfaktorer i de skilda studierna varför de kanske inte är tillräckligt lika för att kunna vägas samman. Man kan då göra en ”forest plot” utan sammanvägning. I vissa fall kan emellertid observationsstudier vara ett alternativ som är gott nog beroende på systematisk brist på randomiserade studier, t ex studier av långsiktiga biverkningar [19].

Komplexiteten kring observationsstudier som inte är likvärdiga vid baslinjen illustreras i Exempel 9.1.

Exempel 9.1 Observationsstudie med skillnader i baslinjedata.

Syftet i en observationsstudie var att undersöka om multidisciplinär vård (MDC) påverkar dödligheten för äldre patienter med kronisk njursjukdom [20]. I en logistisk regression använde man erhållandet av MDC som beroende variabel och ett antal riskfaktorer som beroende variabler. Med stöd av denna modell kunde man därefter räkna fram ett sannolikhetsvärde för att en given patient skulle få MDC. Efter att varje patient fått ett sådant värde (”propensity score”) matchade man patienterna parvis. Därefter jämförde man överlevnadskurvor för dem som fått MDC med dem som inte fått denna vård. Resultatet var att de som fått MDC hade en tydligt lägre momentan risk att dö jämfört med kontrollgruppen med en hasardkvot på 0,50 (95 % KI, 0,35 till 0,71).

För att få en överblick över likheter och olikheter avseende inkluderade observationsstudier kan det krävas att man tabellerar ytterligare information än den som normalt tabelleras för randomiserade studier. Det kan röra sig om vilka variabler som ingår i den modell man använder för att hantera selektionsbias samt själva modellen. Detta exemplifieras i Tabell 9.1 med studier rörande program med multidisciplinära team för sjuka äldre jämfört med standardbehandlingar.

Tabell 9.1 Modell, variabler och matchningsprocedur.

Study	Matching/ adjustment	Variables	Outcome measure
Hemmelgarn et al 2007 [20]	Logistic regression for propensity scores Greedy matching algoritms on propensity scores at ratio 1:1	*Independent:* age, gender, index GFR, diabetes, co-morbidity score, and medication use including angiotensin-converting enzyme, inhibitor or angiotensin receptor blockers, β-blockers, calcium channel blockers, antiarrhythmics, diuretics, cholesterol-lowering agents, and nonsteroidal anti-inflammatory drugs *Dependent:* Assignment to MDC-group	HR 0.50 (0.35, 0.71) In favour of intervention
Wong et al 2006 [21]	Logistic regression for estimating independent risk and adjustment for confounders	*Intervention:* ACE vs other units *Confounders:* age, sex, Apache II score, Charlson's index score, Cumulative Illness Rating Scale score, Geriatric Prognostic Index score, Internal medicine physician service	HR 1.36 (1.10, 1.67) In favour of intervention
Meissner et al 1989 [22]	Adjusted for outliers		WMD 1.80 (–0.85, 4.45) In favour of control
Stewart et al 1999 [23]	No adjustment		WMD –1.10 (–3.83, 1.63) In favour of intervention
Zelada et al 2009 [24]	Logistic regression for estimating independent risk and adjustment for confounders	*Intervention:* Geriatric unit vs usual care unit *Confounders:* age, Mental status score <21, Geriatric depression score >5, Baseline dependency in ≥1 BADL, Apache II score, Comorbidity Charlson index score	OR 4.24 (1.50, 11.99) In favour of intervention

HR = Hazard ratio; OR = Odds ratio; WMD = Weighted mean difference

Om metoden för matchning bedöms vara tillräckligt lika, kan metaanalyser genomföras på samma sätt som för randomiserade studier (förutsatt att de är tillräckligt lika i andra väsentliga avseenden). Om skillnaderna är för stora kan man göra forestplottar, men utan att väga samman effekterna (Figur 9.6).

Metaanalys av diagnostik och psykometri

Ytterligare ett tillämpningsområde för metaanalyser är *diagnostik* och *psykometri* [2]. Detta område är mycket komplext och det pågår för närvarande ett omfattande utvecklingsarbete. Diagnostiska respektive psykometriska studier kan såväl vara randomiserade studier som observationsstudier [25]. Metaanalyserna kan avse olika aspekter av aktuella kliniska processer, t ex hur tillförlitliga testerna är eller patientnyttan av att genomföra testerna vilket är något annat. Testets tillförlitlighet kan jämföras med ett validerat referenstest eller faktiska patienttillstånd. Och vidare, en sammanhängande diagnostisk strategi, som inkluderar flera tester samt behandlingar, kan jämföras med alternativa diagnostiska strategier.

Randomiserade studier kan förekomma t ex om patientnyttan är central och alternativa diagnostiska strategier jämförs. I detta fall kan gängse metoder för metaanalys av resultat från studier av interventionseffekter användas. Försiktighet vid tolkningar av sammanvägda effekter krävs emellertid bl a beroende på de komplexa interventions- och jämförelsealternativen. Om testets tillförlitlighet avseende sensitivitet respektive specificitet är det centrala för översikten krävs andra sammanvägningsmetoder jämfört med sammanvägning av interventionseffekter [26,27]. Det förekommer ytterligare mått som vägs samman från diagnostiska studier, t ex diagnostiska oddskvoter, summerade ROC-kurvor och AUC (”area under curve”). Allt detta beskrivs i detalj i Kapitel 7 om diagnostik.

När det gäller tillförlitlighet avseende *psykometriska tester* förekommer det att man gör metaanalyser avseende korrelationen mellan instrument, t ex ett frågeformulär som fylls i av patienten själv och ett frågeformulär som fylls i av kliniker. Det förekommer åtminstone två metoder för att väga samman dessa korrelationer, en som utvecklats av Hedges och Olkin samt en som Hunter och Schmidt tagit fram. I det första fallet används spridningen vid sammanvägningen och i det andra fallet antalet individer [2,28].

Slutligen bör det nämnas att det utvecklats metaanalysmetoder för att hantera underlag där direkta jämförelser av relevans för praktiken saknas. Man kan t ex vilja veta vilka effekterna skulle bli om två interventioner ställdes mot varandra, men där de båda endast jämförts med ett tredje alternativ. Ett exempel rör trombocythämmare för personer med akuta koronara syndrom med ticagrelor [15] respektive prasugrel [29]. Båda dessa läkemedel har i dagläget jämförts med clopidogrel, men inte med varandra. Genom att

det finns en gemensam nämnare i clopidogrel finns det en möjlighet att försöka skatta effekten av den hypotetiska direkta jämförelsen mellan ticagrelor och prasugrel [30]. Det finns även exempel på försök att skatta liknande hypotetiska effekter där hela nätverk av relaterade resultat vägas samman i så kallade nätverksmetaanalyser [31,32]. SBU:s hållning till dessa typer av metaanalyser är att de kan vara användbara som analytiska verktyg, men att de endast i sällsynta undantagsfall skulle kunna ingå i en evidensprofil (bl a beroende på att de statistiska antaganden som krävs sällan är uppfyllda).

9

Något om programvaror

Det finns flera olika program som kan användas för metaanalys. Det enklaste programmet, som för närvarande är fritt tillgängligt på internet, är Review Manager (RevMan) som tagits fram inom Cochrane Collaboration (www.cochrane.org). Programmet följer internationellt etablerade konventioner, men klarar ännu inte t ex hasardkvoter och mer komplicerade former av metaanalys som t ex metaregression, nätverksmetaanalys, diagnostiska metaanalyser, metaanalys av korrelationer. Comprehensive Meta-Analysis (CMA) har några fler funktioner än RevMan (t ex metaregression), men är avgiftsbelagd (www.meta-analysis.com). Meta-DiSc som utvecklats speciellt för metaanalyser inom diagnostik är ännu fritt tillgängligt via internet (www.hrc.es/investigacion/metadisc_en.htm). Det program som har mest möjligheter, men som är avgiftsbelagt och kräver mest erfarenhet är det generella statistikprogrammet Stata (www.stata.com). Sedan bör man inte glömma att mycket går att genomföra i Excel som finns inom Officepaketet.

Ett område i snabb utveckling

Metaanalyser och relaterade metoder är föremål för en snabb utveckling. Gamla arbetssätt modifieras och nya metoder tillkommer. Metaanalysen har kommit längst när det handlar om interventionseffekter, men inte lika långt när det t ex handlar om diagnostik. I dessa sammanhang är det av stor vikt att följa utveckling via internationella HTA-nätverk såsom Cochrane Collaboration, GRADE Working Group m fl inom vilka konventioner, systematik och transparens utvecklas. Av speciell betydelse för arbetet med metaanalyser är PRISMA statement (en vidareutveckling av QUORUM statement). Tre förändringar det senaste decenniet kan lyftas fram: (a) fokus har förskjutits från enskilda studier till effektmått (vilka kan inkludera resultat från flera studier) då risk för bias bedöms, (b) betydelsen av kontext och extern validitet har betonats mer än tidigare, och (c) gamla former av evidenshierarkier har börjat problematiseras (vilket innebär att det inte är omöjligt att resultat från observationsstudier kan bedömas ha låg risk för bias).

Referenser

1. Haynes RB, Devereaux PJ, Guyatt GH. Clinical expertise in the era of evidence-based medicine and patient choice. ACP J Club 2002;136:A11-4.

2. Borenstein M, Hedges LV, Higgins JPT, et al. Introduction to meta-analysis. Chichester: John Wiley & Sons Ltd; 2009.

3. Bond GR, Witheridge TF, Dincin J, Wasmer D. Assertive community treatment for frequent users of psychiatric hospitals in a large city: A controlled study. Am J Community Psychol 1990;18:865-91.

4. Clarke GN, Herinckx HA, Kinney RF, Paulson RI, Cutler DL, Lewis K, Oxman E. Psychiatric hospitalizations, arrests, emergency room visits, and homelessness of clients with serious and persistent mental illness: findings from a randomized trial of two ACT programs vs. usual care. Ment Health Serv Res 2000;2:155-64.

5. Conrad KJ, Hultman CI, Pope AR, Lyons JS, Baxter WC, Daghestani AN, et al. Case managed residential care for homeless addicted veterans: Results of a true experiment. Medical Care 1998;36: 40-53.

6. Cox GB, Walker RD, Freng SA, Short BA, Meijer L, Gilchrist L. Outcome of a controlled trial of the effectiveness of intensive case management for chronic public inebriates. J Stud Alcohol 1998;59:523-32.

7. Lehman AF, Dixon LB, Kernan E, DeForge BR, Postrado LT. A randomized trial of assertive community treatment for homeless persons with severe mental illness. Arch Gen Psychiatry 1997;54:1038-43.

8. Rosenheck R, Kasprow W, Frisman L, Liu-Mares W. Cost-effectiveness of supported housing for homeless persons with mental illness. Arch Gen Psychiatry 2003;60:940-51.

9. Lipton FR, Nutt S, Sabatini A. Housing the homeless mentally ill: a longitudinal study of a treatment approach. Hosp Community Psychiatry 1988;39:40-5.

10. Morse GA, Calsyn RJ, Allen G, Tempelhoff B, Smith R. Experimental comparison of the effects of three treatment programs for homeless mentally ill people. Hosp Community Psychiatry 1992;43:1005-10.

11. Morse GA, Calsyn RJ, Dean Klinkenberg W, Helminiak TW, Wolff N, Drake RE, et al. Treating homeless clients with severe mental illness and substance use disorders: costs and outcomes. Community Ment Health J 2006;42:377-404.

12. Sosin MR, Bruni M, Reidy M. Paths and impacts in the progressive independence model: a homelessness and substance abuse intervention in Chicago. J Addict Dis 1995; 14:1-20.

13. Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions Version 5.1.0. The Cochrane Collaboration. Available from www.cochrane-handbook.org; 2008.

14. Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database Syst Rev 2009:MR000006.

15. Wallentin L, Becker RC, Budaj A, Cannon CP, Emanuelsson H, Held C, et al. Ticagrelor versus clopidogrel in patients with acute coronary syndromes. N Engl J Med 2009;361:1045-57.

16. Cannon CP, Husted S, Harrington RA, Scirica BM, Emanuelsson H, Peters G, et al. Safety, tolerability, and initial efficacy of AZD6140, the first reversible oral adenosine diphosphate receptor antagonist, compared with clopidogrel, in patients with non-ST-segment elevation acute coronary syndrome: primary results of the DISPERSE-2 trial. J Am Coll Cardiol 2007;50:1844-51.

17. Campbell DS, Stanley JC. Experimental and quasi-experimental designs for research. Chicago: Rand McNally & Company; 1963.

18. Shadish WC, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference. Boston/New York: Houghton Mifflin Company; 2002.

19. Golder S, Loke YK, Bland M. Meta-analyses of adverse effects data derived from randomised controlled trials as compared to observational studies: methodological overview. PLoS Med 2011;8:e1001026.

20. Hemmelgarn BR, Manns BJ, Zhang J, Tonelli M, Klarenbach S, Walsh M, Culleton BF. Association between multidisciplinary care and survival for elderly patients with chronic kidney disease. J Am Soc Nephrol 2007;18:993-9.

21. Wong RY, Chittock DR, McLean N, Wilbur K. Discharge outcomes of older medical in-patients a specialized acute care for elders unit compared with non-specialized units. Canadian Journal of Geriatrics 2006;9:96-101.

22. Meissner P, Andolsek K, Mears P, Fletcher B. Maximizing the functional status of geriatric patients in an acute community hospital setting. Gerontologist 1989;29:524-8.

23. Stewart M, Suchak N, Scheve A, Popat-Thakkar V, David E, Laquinte J, Gloth FM 3rd. The impact of a geriatrics evaluation and management unit compared to standard care in a community teaching hospital. Md Med J 1999;48:62-7.

24. Zelada MA, Salinas R, Baztán JJ. Reduction of functional deterioration during hospitalization in an acute geriatric unit. Arch Gerontol Geriatr 2009;48:35-9.

25. Brozek JL, Akl EA, Jaeschke R, Lang DM, Bossuyt P, Glasziou P, et al. Grading quality of evidence and strength of recommendations in clinical practice guidelines: Part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies. Allergy 2009;64:1109-16.

26. Gatsonis C, Paliwal P. Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer. AJR Am J Roentgenol 2006;187:271-81.

27. Zamora J, Abraira V, Muriel A, Khan K, Coomarasamy A. Meta-DiSc: a software for meta-analysis of test accuracy data. BMC Med Res Methodol 2006;6:31.

28. Field AP. Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed- and random-effects methods. Psychol Methods 2001;6:161-80.

29. Wiviott SD, Braunwald E, McCabe CH, Montalescot G, Ruzyllo W, Gottlieb S, et al. Prasugrel versus clopidogrel in patients with acute coronary syndromes. N Engl J Med 2007;357:2001-15.

30. Biondi-Zoccai G, Lotrionte M, Agostoni P, Abbate A, Romagnoli E, Sangiorgi G, et al. Adjusted indirect comparison meta-analysis of prasugrel versus ticagrelor for patients with acute coronary syndromes. Int J Cardiol 2011;150:325-31.

31. Woods BS, Hawkins N, Scott DA. Network meta-analysis on the log-hazard scale, combining count and hazard ratio statistics accounting for multi-arm trials: a tutorial. BMC Med Res Methodol 2010;10:54.

32. Wandel S, Jüni P, Tendal B, Nüesch E, Villiger PM, Welton NJ, et al. Effects of glucosamine, chondroitin, or placebo in patients with osteoarthritis of hip or knee: network meta-analysis. BMJ 2010;341:c4675.

10. **Evidensgradering**

VERSION 2011:1.4

Det sista steget i utvärderingen är att bedöma hur starkt det samlade vetenskapliga underlaget är. SBU använder det internationellt utarbetade GRADE-systemet [1,2] för att klassificera styrkan på det vetenskapliga underlaget.

GRADE är ett system som utvecklas kontinuerligt via GRADE Working Group där även SBU ingår. Bakgrunden till att GRADE bildades är att det finns en flora av olika system som används parallellt för att gradera evidens och styrka på rekommendationer. Denna mångfald har lett till en viss förvirring, och att många har upplevt att viktiga steg i tidigare utvärderingsrapporter ibland utelämnats eller varit otydliga.

I princip bygger GRADE på erfarenheter från andra graderingssystem men med en tydligare fokus på risk–nyttaperspektivet. GRADE har redan anammats av internationella aktörer som t ex WHO, NICE, Cochrane Collaboration och BMJ. I Sverige används GRADE av bl a SBU och Socialstyrelsen.

GRADE:s evidensgradering bygger på en fyrgradig skala från starkt, måttligt och lågt till mycket lågt vetenskapligt underlag. SBU har valt att ersätta ordet ”lågt” med ”begränsat” och ”mycket lågt” med ”otillräckligt”, men de innebär i princip samma sak. Skälen till att vi ändrat ordvalet är att det överensstämmer bättre med SBU:s tidigare nomenklatur och att vi anser att det ger lite mer vägledning. Evidensstyrkorna blir då med SBU:s terminologi starkt (⊕⊕⊕⊕), måttligt (⊕⊕⊕○), begränsat (⊕⊕○○) och otillräckligt (⊕○○○) vetenskapligt underlag. Begränsat underlag kan vara tillräckligt för att tillämpa metoden i klinisk praxis om andra kriterier är uppfyllda, t ex rimlig kostnadseffektivitet. Otillräckligt underlag tydliggör att vi måste ha mer forskning innan metoden kan tillämpas i stor skala. Förenklat kan man säga att ett starkt vetenskapligt underlag är så stabilt att det är liten risk för att ny forskning kommer att komma fram till nya slutsatser. På samma sätt innebär ett begränsat vetenskapligt underlag att det är högre risk för att nya studier kan förändra slutsatsen.

Detta avsnitt beskriver tillvägagångssättet för att använda GRADE. Observera att systemet hittills är utvecklat för interventions-/behandlingsstudier. Det kan dock på likartat sätt användas vid studier av orsakssamband. Bedömning av det vetenskapliga underlaget för diagnostiska studier och för studier med kvalitativ metodik med hjälp av GRADE är fortfarande i en utvecklingsfas.

I GRADE-systemet finns också en rekommendationsdel som inte används av SBU.

En sammanfattande resultattabell ger en bra och kortfattad bild av underlaget

Ett lämpligt första steg är att göra en sammanfattande resultattabell. Detta förenklar det fortsatta arbetet. Ett exempel på en sådan resultattabell syns i Tabell 10.1. I tabellen bör den sammanvägda effekten för samtliga viktiga effektmått redovisas separat (t ex dödlighet, funktion och livskvalitet). I vissa fall förtecknas även effekter som mer är att betrakta som surrogatmått (t ex HbA_{1c} och blodtryck). Effektmåtten innefattar positiva effekter, men även negativa som biverkningar och komplikationer bör tas med. De olika effektmåtten bör tabelleras hierarkiskt så att de viktigaste står före de mindre viktiga. Som framgår av Tabell 10.1 fyller man i kolumnerna ”Vetenskapligt underlag” och ”Kommentarer” i ett senare skede efter den samlade bedömningen.

Tabell 10.1 Sammanfattande resultattabell (”Summary of findings”).
Effekten av antibiotikaprofylax jämfört med placebo vid käkkirurgiska ingrepp.

Effektmått	Studiedesign Antal patienter (antal studier)	Medelrisk i standard-grupp (min–max)	Relativ risk (95% KI)	Absolut effekt per 1 000 patienter	Veten-skapligt underlag	Kom-men-tarer
Sårinfek-tioner vid operation av käkfrakturer	RCT 461 (3)	39% (20–62%)	RR 0,25 (0,15 till 0,41)	259 färre		

I exemplet med sårinfektioner vid operationer av käkfrakturer (Tabell 10.1) fanns tre RCT med sammanlagt 461 patienter där det gick att göra en metaanalys och på så sätt få fram en sammanvägd effekt i numeriska termer [3]. I många fall går det dock inte att få fram så preciserade data på medelvärden och riskdifferens. När data är för heterogena är det inte möjligt att väga ihop resultaten i en poolad metaanalys med en absolut eller relativ riskdifferens. Effekten kan då istället redovisas som min–maxvariation för studierna.

Om det finns gott om välgjorda randomiserade kontrollerade studier (RCT) behöver normalt sett inte observationsstudier inkluderas i bedömningen av positiva effekter. Om det inte finns randomiserade studier eller om de randomiserade studierna är bristfälliga eller har för kort uppföljningstid kan observationsstudier ge viktig tilläggsinformation och bidra till den samlade evidensgraderingen i både positiv och negativ riktning. RCT och observationsstudier redovisas dock separat i resultattabellen. Vid bedömning av risker är det ofta viktigt att inkludera observationsstudier eftersom RCT i de flesta fall inte är utformade för att besvara frågor om långsiktiga risker.

Preliminär evidensstyrka

Som nämnts ovan sätts en evidensstyrka för varje utfallsmått som finns med i resultattabellen.

I arbetet med GRADE utgår man från en preliminär evidensstyrka. Den baseras enbart på vilket studieupplägg (studiedesign) som studierna som ingår i det vetenskapliga underlaget består av. Den preliminära evidensstyrkan justeras sedan uppåt eller neråt beroende på ett antal kvalitetsfaktorer som beskrivs närmare nedan. Om underlaget till största delen består av randomiserade studier där risken för systematiska fel är minst så bedöms det preliminärt som starkt.

Faktaruta 10.1 Preliminär evidensstyrka baserad på studiedesign och skäl för ned- eller uppgradering av evidensgraderingen.

Preliminär evidensstyrka för interventionsstudier:

Evidensstyrka	Studiedesign
Stark (⊕⊕⊕⊕)	Randomiserade studier
Måttligt stark (⊕⊕⊕◯)	
Begränsad (⊕⊕◯◯)	Observationsstudier; kohort- och fall–kontrollstudier
Otillräcklig (⊕◯◯◯)	Fallstudier m m

Sedan kan evidensstyrkan sänkas eller höjas enligt nedanstående:

Sänk gradering om	Höj gradering om
• Brister i studiekvalitet (maximalt –2) • Bristande överensstämmelse mellan studierna (maximalt –2) • Brister i överförbarhet/relevans (maximalt –2) • Bristande precision (maximalt –1) • Hög sannolikhet för publikationsbias (maximalt –2)	• Stora effekter och inga sannolika ”confounders” (maximalt +2) • Tydligt dos–responssamband (maximalt +1) • ”Confounders” som inte är med i analysen borde leda till bättre behandlingsresultat i kontrollgruppen, dvs hög sannolikhet att effekten underskattas (maximalt +1)

När det gäller diagnostiska tillförlitlighetsstudier (”accuracy”) har GRADE-gruppen tagit ställning för att den preliminära evidensstyrkan ska utgå från stark även för observationsstudier [4]. Detta kan diskuteras, men SBU accepterar denna utgångspunkt tills vidare. Däremot är det mycket viktigt att man analyserar om förbättrad diagnostik i slutändan ger patientnytta.

Sju faktorer påverkar den slutliga evidensstyrkan

För att fastställa den slutliga evidensstyrkan bedömer expertgruppen hur tillförlitligt det vetenskapliga underlaget är. Den preliminära evidensstyrkan *sänks* om underlaget är osäkert med avseende på:

10

- studiekvalitet
- samstämmighet/överensstämmelse
- överförbarhet/relevans
- precision i data
- risk för publikationsbias.

Evidensstyrkan kan dras ner med ett eller två steg för varje faktor beroende på hur stora bristerna är. Om bristerna är mindre allvarliga kan man notera det utan att gradera ner. Om det finns mindre allvarliga brister med avseende på flera av faktorerna kan det leda till att evidensstyrkan totalt sett dras ner ett steg. Man bör dock komma ihåg att observationsstudier redan i utgångsläget graderats ned pga studiedesign och därför inte generellt bör graderas ned ytterligare pga brist på ”confounding”-kontroll. Vid allvarliga brister i ”confounding”-kontrollen kan det dock vara motiverat att även nedgradera observationsstudier. Däremot kan brister som rör samstämmighet, överförbarhet, precision och risk för publikationsbias motivera nedgradering av observationsstudier.

I vissa fall finns det också skäl till att *öka* evidensstyrkan med ett eller två steg. Detta gäller när det vetenskapliga underlaget består av stora, välgjorda observationsstudier med god kontroll för förväxlingsfaktorer. De tre faktorer som kan höja evidensstyrkan är:

- stora effekter
- dos–responssamband
- hög sannolikhet att effekten i studien är underskattad.

Studiekvalitet

Under granskningsfasen bedömdes varje studies kvalitet individuellt. I detta steg görs en samlad värdering av hela materialet. I första hand vägs de traditionella faktorerna som t ex randomisering, blindning och bortfall in. Hur välgjorda är studierna med avseende på randomisering totalt sett? Är några mycket välgjorda medan resterande studier har ett oklart randomiseringsförfarande? Även andra faktorer som är belysta i granskningsmallarna kan vara viktiga att beakta, likaså ämnesspecifika metodproblem som experterna har identifierat.

För kohortstudier och andra observationsstudier är frågan om jämförbarhet mellan försöks- och kontrollgrupp central. Det innebär att bedömningen av den sammanvägda studiekvaliteten i hög grad beror på i vad mån studierna har kontrollerat för förväxlingsfaktorer (”confounders”, se vidare i Kapitel 6).

Exempel 10.1 Risk för att effekten överskattats.

Observationsstudier visade att östrogenbehandling minskade risken för hjärt- och kärlsjukdom. Eftersom man visste att östrogenbehandling var vanligare bland kvinnor från högre socialgrupper kunde man ha misstänkt att den observerade effekten var överskattad och därigenom nedgraderat tilltron till sambandet.

Diagnostiska studier kan graderas ner om rekryteringen inte är konsekutiv, om utvärderarna inte är blindade och om studien har någon form av verifikationsbias.

GRADE-klassificeringen innebär att evidensstyrkan kan justeras ned ett steg om det finns allvarliga kvalitetsbrister i underlaget och två steg om begränsningarna är mycket allvarliga. Observera att GRADE i sig inte kräver att underlaget ska ha minst medelhög kvalitet. För SBU:s vidkommande, där studier med låg kvalitet sorterats bort, är det i praktiken mycket ovanligt att evidensstyrkan dras ner två steg pga metodologiska brister.

Samstämmighet/överensstämmelse

Här bedöms i vilken utsträckning studierna kommer fram till samma resultat. Pekar de åt samma håll och är effektstorleken av jämförbar storlek i de olika studierna? Meta-analyser kan vara en god hjälp för att bedöma graden av samstämmighet.

Samstämmigheten kommer att vara beroende av hur likartade studierna är med avseende på population, exakt hur interventionen genomfördes och vilken kontrollgrupp som användes. En annan viktig faktor är om större delen av studierna har genomförts av samma forskargrupp.

Generellt sett ökar trovärdigheten i det samlade materialet om studierna har gjorts av olika forskargrupper med olika populationer och studierna samstämmigt pekar i samma riktning. Om studierna visar såväl över- som underrisker kan evidensstyrkan minskas med ett steg. Detsamma gäller om effektstorlekarna varierar kraftigt mellan studierna, vilket leder till en ökad osäkerhet.

I vissa fall kan skillnaderna förklaras med olikheter i studierna, t ex att de undersökt olika populationer. I sådana fall kan det vara mer lämpligt att dela upp materialet och formulera slutsatser för de olika populationerna var för sig.

Överförbarhet/relevans

Med överförbarhet menas i vilken utsträckning det vetenskapliga underlaget är generaliserbart och relevant för svenska förhållanden. Viktiga frågor att överväga är t ex hur väl populationen överensstämmer med den man ser i daglig svensk praxis, om interventionen utförs på samma sätt som i Sverige, om kontrollgruppen är relevant och om sjukvårdsmiljön är rimligt likartad.

Ett exempel på bristande överförbarhet är att kontrollgruppen får en behandling som inte finns tillgänglig i Sverige. Det går då inte att avgöra hur effektiv interventionen är jämfört med sedvanlig behandling i svensk praxis.

Om överförbarheten och relevansen är bristfällig kan evidensstyrkan justeras ned ett eller två steg. På samma sätt som när det gäller studiekvalitet har sannolikt studier som är mindre relevanta för svenska förhållanden redan sorterats bort (med hjälp av relevansmallen, se Kapitel 5 och Bilaga 1).

Ett specialfall är när det bara finns en studie som mäter effekten med hjälp av ett visst utfallsmått. SBU:s bedömning är att överförbarheten vanligtvis kräver minst två studier. Det innebär att evidensstyrkan i detta fall oftast justeras ner ett steg. Undantag kan t ex vara när underlaget består av en stor mycket välgjord randomiserad multicenterstudie.

Precision i data

Detta kriterium uppskattar hur osäker den sammanvägda effekten är. Få observationer och breda konfidensintervall i de olika studierna kommer att leda till sämre precision. Precisionen beror av antalet händelser, antal personer i grupperna och den relativa riskminskningen.

Ett hjälpmedel för att bedöma om precisionen är osäker är att göra en poweranalys baserad på det totala antalet observationer i de inkluderade studierna. Om antalet observationer i dessa inkluderade studier är mindre än det antal som krävs för att visa statistiskt signifikanta resultat kan det finnas anledning att nedgradera för bristande precision. Om studierna är mycket små bör man även om utfallen är statistiskt signifikanta vara uppmärksam på om skillnaderna i ingångsvärden (baslinjedata) är stora. Skiljer sig baslinjedata mycket åt mellan grupperna kan det finnas anledning att nedgradera för precision i data.

Risk för publikationsbias

Med publikationsbias avses att delar av det vetenskapliga underlaget inte är tillgängligt i publicerade studier. Risken för publikationsbias ökar bl a om underlaget enbart består av små studier som är utförda av samma forskargrupp och som har stora metodbrister. Tidiga utvärderingar av nya metoder faller ofta inom den här kategorin. Analoga situationer är när underlaget utgörs av företagssponsrade studier av t ex ett läkemedel. Om samtliga studier av en ny metod har innovatören som huvudförfattare finns det också skäl att överväga publikationsbias.

Effekterna av selektiv publicering framgår väl i en svensk studie från Läkemedelsverket [6]. Systematiska översikter [7] och många andra studier [8,9] pekar entydigt på att studier som sponsrats av industrin eller andra aktörer med egenintressen av resultaten överskattar effekterna av sina produkter. Kostnadseffektanalyser som utförts av läkemedelsföretag visade t ex mer än två gånger så ofta kostnadseffektkvoter under 20 000 amerikanska dollar per kvalitetsjusterat vunnet levnadsår som icke-industrisponsrade studier [8,9].

Det är ofta svårt att avgöra om det finns publikationsbias, men det finns några metoder som underlättar bedömningen. Ett sätt är att använda centrala register (t ex www.controlled-trials.com och www.clinicaltrials.gov) över påbörjade kliniska prövningar. En kontroll av vilka studier som finns i registren bör ingå i bedömningen.

Registren har funnits sedan början av 2000-talet och kan ge värdefull information om studier som startats det senaste decenniet. Studier som, enligt registret, förefaller vara avslutade sedan flera år tillbaka kan vara en möjlig källa till publikationsbias. Omvänt kan stora pågående studier leda till en ökad risk för att evidensstyrkan påverkas av deras resultat i framtiden, så även för bedömning av osäkerhet kan registren ge viss vägledning.

Ytterligare ett sätt att påvisa publikationsbias är att titta i databaser för konferensabstrakt. Förekomst av abstrakt för studier som inte publicerats inom ett rimligt antal år är ett starkt stöd för att publikationsbias föreligger. Detta är dock ett arbetskrävande moment.

Om det finns många studier kan risken bedömas med hjälp av en så kallad ”funnel plot” (Bilaga 9). Som en tumregel krävs minst fem studier för att analysen ska vara meningsfull.

Effektstorlek

Detta kriterium kan användas om det finns minst två stora, välgjorda observationsstudier med god kontroll för förväxlingsfaktorer. Om effekten är hög ökar sannolikheten för att det funna sambandet är kausalt.

GRADE anger att den samlade evidensstyrkan kan höjas med ett steg om den sammanvägda effekten definierad som relativ risk (RR) är större än 2 (RR >2,0 alternativt RR <0,5). Om RR >5,0 (alternativt RR <0,2) kan evidensstyrkan höjas med två steg. Om man tidigare bedömt att studiekvaliteten är bristfällig kan man avstå från att höja evidensgraderingen pga stor effektstorlek.

10

Observera att om utfallsmåttet är oddskvoter (OR) kan effekterna vara överskattade om utfallen är vanliga (>10 procent). Gränserna kan då behöva justeras. Se mer i statistikavsnittet (Bilaga 9) där skillnaden mellan RR och OR förklaras tillsammans med en formel för hur man kan omvandla OR till RR och vice versa.

Exempel 10.2 Stor effekt kan höja evidensstyrkan.

En metaanalys av observationsstudier visade att cykelhjälmar reducerade risken för huvudskador med en oddskvot på 0,31 (0,26–0,37) [5]. Detta är en stor effekt och leder till att evidensstyrkan justeras upp ett steg. Om det inte finns någon anledning till att justera ner evidensstyrkan pga brister i underlaget skulle det medföra att det vetenskapliga underlaget blir måttligt starkt (⊕⊕⊕◯).

Dos–responssamband

Även detta kriterium är begränsat till att gälla stora, välgjorda observationsstudier. Dos–respons kan avse både effekter och risker. Ett dos–responssamband ökar trovärdigheten för att en åtgärd har effekt. Dos–respons kan gälla för både läkemedel och andra insatser. Effekten kan mätas dels inom en studie och dels mellan studier. Generellt är ett dos–responssamband mycket mer trovärdigt när det visats inom en studie än när det har visats genom jämförelser mellan studier.

Ett dos–responssamband kan öka evidensstyrkan med ett steg.

Exempel 10.3 Dos–responssamband kan höja evidensstyrkan.

I SBU-rapporten ”Mat vid diabetes” [10] studerades bl a risken för hjärtinfarkt bland diabetiker med varierande alkoholkonsumtion. I tre stora observationsstudier med totalt 10 312 patienter såg man i samtliga studier ett dos–responsförhållande så att gruppen med högre alkoholkonsumtion hade betydligt lägre relativ risk än de med låg eller ingen konsumtion. Expertgruppen uppgraderade därför evidensstyrkan till måttligt starkt vetenskapligt underlag pga dos–responsförhållande.

Hög sannolikhet att effekten i studien är underskattad

Vid enstaka tillfällen kan evidensstyrkan justeras upp om det är mycket sannolikt att studierna underskattat effekten. Detta kan gälla om de förväxlingsfaktorer (”confounders”) som studien inte har kunnat justera för, talar för att effekten är underskattad.

Exempel 10.4 Underskattad effekt kan höja evidensstyrkan.

I en systematisk översikt som omfattade 38 miljoner patienter var dödstalen högre på vinstdrivande privata sjukhus än på icke-vinstdrivande privata sjukhus [11]. GRADE Working Group menar att det är sannolikt att patienterna var sjukare på de icke-vinstdrivande sjukhusen och att de vinstdrivande sjukhusen hade större resurser och fler patienter som var väl försäkrade. Evidensen för att dödstalen verkligen är högre på vinstdrivande sjukhus har då stärkts.

Samlad evidensgradering

Det är viktigt att projektgruppen kommenterar skriftligt hur underlaget påverkats av var och en av ovan diskuterade faktorer, som motivering till bedömningen. För att underlätta arbetet med evidensgradering kan man göra en evidenstabell (Tabell 10.2) I denna noteras den preliminära evidensstyrkan och de justeringar som görs i nästa steg.

När det gäller exemplet antibiotikaprofylax vid käkkirurgi med utfallsmåttet sårinfektion kan man i första steget göra en evidenstabell (Tabell 10.2) som underlag för en sammanfattande resultattabell (Tabell 10.3).

I detta fall bedömdes brister i studiekvaliteten motivera en nedgradering liksom de sammanlagda bristerna av att studierna hade få utfall och använt olika antibiotika.

Tabell 10.2 Evidenstabell. Sårinfektioner, antibiotikaprofylax jämfört med placebo vid operation av käkfrakturer.

Studier Patienter	Design	Studie-kvalitet	Överens-stämmelse	Överför-barhet	Oprecisa data	Publika-tionsbias	Effektstorlek	Dos–respons	Förväxlings-faktor
3 461	RCT ⊕⊕⊕⊕	–1*	0	0**	–1***	0	0	0	0

** Ingen blindning, en studie har ingen uppgift om bortfall*

*** Olika antibiotika*

**** Få utfall, tillsammans med ** –1*

Den sammanfattande resultattabellen för antibiotikaprofylax redovisas därefter i Tabell 10.3. Den relativa risken var här stor och kunde ha motiverat en uppgradering, men expertgruppen bedömde att bristande studiekvalitet i detta fall inte motiverade en uppgradering.

När det vetenskapliga underlaget för samtliga utfallsmått har bestämts ger resultattabellen en samlad bild av kunskapsläget för en viss fråga.

Tabell 10.3 Sammanfattande resultattabell (”Summary of findings”).
Effekten av antibiotikaprofylax jämfört med placebo vid käkkirurgiska ingrepp.

Effektmått	Studie-design Antal patienter (antal studier)	Medelrisk i standard-grupp (min–max)	Relativ risk (95% KI)	Absolut effekt per 1 000 patienter	Veten-skapligt underlag	Kommen-tarer
Sårinfek-tioner vid operation av käkfrakturer	RCT 461 (3)	39% (20 till 62%)	RR 0,25 (0,15 till 0,41)	259 färre	⊕⊕◯◯ Begränsat	Svagheter i studie-kvalitet (–1) Få utfall (–1)

Tolkning av evidensstyrkan

Även om utvärderingen inte mynnar ut i en rekommendation så ger den vägledning för hälso- och sjukvården. Om det vetenskapliga underlaget är otillräckligt indikerar det behov av mer forskning innan metoden kan ingå i rutinsjukvård. En begränsad evidensstyrka kan motivera att metoden används i hälso- och sjukvården under förutsättning att den uppfyller andra krav på acceptabel balans mellan risk och nytta, kostnadseffektivitet och att den är etiskt acceptabel. Vid måttlig eller hög evidensstyrka är det vetenskapliga underlaget gott och motiverar sannolikt att metoden tillämpas under förutsättning att den ekonomiska, etiska och sociala analysen i utvärderingen ger stöd för metoden.

Faktaruta 10.2 Studiekvalitet, evidensstyrka och slutsatser.

Studiekvalitet avser den vetenskapliga kvaliteten hos en enskild studie och dess förmåga att besvara en viss fråga på ett tillförlitligt sätt.

Evidensstyrkan är en bedömning av hur starkt det sammanlagda vetenskapliga underlaget är för att besvara en viss fråga på ett tillförlitligt sätt. SBU tillämpar det internationellt utarbetade evidensgraderingssystemet GRADE. För varje effektmått utgår man i den sammanlagda bedömningen från studiernas design. Därefter kan evidensstyrkan påverkas av förekomsten av försvagande eller förstärkande faktorer som studiekvalitet, samstämmighet, överförbarhet, effektstorlek, precision i data, risk för publikationsbias och andra aspekter, t ex dos–responssamband.

Evidensstyrka graderas i fyra nivåer:

Starkt vetenskapligt underlag (⊕⊕⊕⊕)
Bygger på studier med hög eller medelhög kvalitet utan försvagande faktorer vid en samlad bedömning.

Måttligt starkt vetenskapligt underlag (⊕⊕⊕○)
Bygger på studier med hög eller medelhög kvalitet med förekomst av enstaka försvagande faktorer vid en samlad bedömning.

Begränsat vetenskapligt underlag (⊕⊕○○)
Bygger på studier med hög eller medelhög kvalitet med försvagande faktorer vid en samlad bedömning.

Otillräckligt vetenskapligt underlag (⊕○○○)
När vetenskapligt underlag saknas, tillgängliga studier har låg kvalitet eller där studier av likartad kvalitet visar motsägande resultat, anges det vetenskapliga underlaget som otillräckligt.

Ju starkare evidens desto mindre sannolikt är det att redovisade resultat kommer att påverkas av nya forskningsrön inom överblickbar framtid.

Slutsatser
I SBU:s slutsatser görs en sammanfattande bedömning av nytta, risker och kostnadseffektivitet.

10

Referenser

1. Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al. Grading quality of evidence and strength of recommendations. BMJ 2004;328:1490.

2. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines 1. Introduction – GRADE evidence profiles and summary of finding tables. J Clin Epidemiol 2011;64:383-94.

3. SBU. Antibiotikaprofylax vid kirurgiska ingrepp. En systematisk litteraturöversikt. Stockholm: Statens beredning för medicinsk utvärdering (SBU); 2010. SBU-rapport nr 200. ISBN 978-91-85413-36-2.

4. Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. GRADE: grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008;336:1106-10.

5. Thompson DC, Rivara FP, Thompson R. Helmets for preventing head and facial injuries in bicyclists. Cochrane Database of Systematic Review 2000, issue 2. Art.nr: CD001855.

6. Melander H, Ahlqvist-Rastad J, Meijer G, Beermann B. Evidence b(i)ased medicine-selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications. BMJ 2003;326:1171-3.

7. Lexchin J, Bero LA, Djulbegovic B, Clark O. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. BMJ 2003;326:1167-70.

8. Jörgensen AW, Maric KL, Tendal B, Faurschou A, Götzche PC. Industry-supported meta-analysis compared with meta-analysis with non-profit or no support: differences in methodological quality and conclusions. BMC Med Res Methodol 2008;8:60.

9. Bell CM, Urbach DR, Ray JG, Bayoumi A, Rosen AB, Greenberg D, Neumann PJ. Bias in published cost effectiveness studies: systematic review. BMJ 2006;332:699-703.

10. SBU. Mat vid diabetes. En systematisk litteraturöversikt. Stockholm: Statens beredning för medicinsk utvärdering (SBU); 2010. SBU-rapport nr 201. ISBN 978-91-85413-37-9.

11. Devereaux PJ, Choi PT, Lacchetti C, Weaver B, Schünemann HJ, Haines T, et al. A systematic review and meta-analysis of studies comparing mortality rates of private for-profit and private not-for-profit hospitals. CMAJ 2002;166:1399-406.

11. Hälsoekonomiska utvärderingar

VERSION 2012:1.1

11

Inledning

Utvärdering av medicinsk teknologi (”health technology assessment”, HTA) definieras som en tvärvetenskaplig analys av hälso- och sjukvårdsmetoder [1]. Utöver att studera de medicinska, sociala och etiska aspekterna av att utveckla, sprida och använda en metod inom hälso- och sjukvården är ekonomiska aspekter också en viktig del av analysen. Att ekonomi har en så pass viktig roll beror på att hälso- och sjukvårdens resurser är begränsade. I kombination med att efterfrågan på hälso- och sjukvård hos befolkningen är hög och dessutom ökar [2,3] uppstår ett gap mellan vad samhället kan erbjuda och vad som efterfrågas av dess invånare. Prioriteringar behöver därför göras när det ska bestämmas vilka behandlingar eller diagnostiska metoder som resurserna ska läggas på. Hälsoekonomiska utvärderingar, som på ett strukturerat sätt jämför kostnader och effekter för två eller flera alternativ, är ett hjälpmedel för att avgöra huruvida en metod ger patienterna så pass mycket hälsa att det står i proportion till vad den kostar.

Analys av ekonomiska aspekter utgör en viktig del av de projekt som SBU genomför. I följande kapitel beskrivs arbetet med att sammanställa och kvalitetsgranska den hälsoekonomiska litteraturen och hur SBU arbetar med egna ekonomiska utvärderingar. Därefter beskrivs de hälsoekonomiska begrepp och analysformer som utgör grunden i arbetet med hälsoekonomiska utvärderingar.

SBU:s arbete med hälsoekonomiska utvärderingar

Hälsoekonomiska aspekter i SBU:s projekt beaktas vanligtvis i en eller flera av följande former:

- sjukdomskostnadskalkyler (”cost-of-illness”)
- hälsoekonomiska utvärderingar
  - systematiska översikter av befintlig litteratur om kostnadseffektivitet (empiriska studier och modeller)
  - egna kostnadseffektanalyser
- budgetpåverkansanalyser (”budget impact analysis”).

Sjukdomskostnadskalkyler (”cost-of-illness”)

Ohälsa kan beskrivas och mätas utifrån olika perspektiv: individens/patientens egen upplevelse (självrapporterad sjuklighet, ”illness”) och/eller utifrån sjukvårdspersonalens definition baserad på medicinska kriterier (diagnostiserad sjuklighet, ”disease”). För en

övergripande beskrivning av sjuklighet och av förändringar i sjukdomspanoramat kan man beräkna de samlade kostnaderna i samhället pga sjukdom och skador. Denna typ av studier brukar kallas för ”cost-of-illness”-studier (COI) [4,5]. Samhällskostnaden för olika sjukdomar ger viss information om problemets storlek men ger dock inte något besked om olika metoders kostnadseffektivitet, och utgör därmed inget beslutsunderlag för fördelning av resurser i sjukvården [6,7]. Sådana beräkningar kan därför lämpligen redovisas i rapportens inledningskapitel.

11

Hälsoekonomiska utvärderingar

Systematiska översikter av hälsoekonomiska utvärderingar

Första steget i analysen av kostnadseffektivitet är att göra en systematisk översikt av den litteratur som är publicerad. En litteratursökning görs då utifrån de söktermer som använts i sökningen av den medicinska litteraturen men kompletterat med ekonomiska sökord. Studiernas relevans bedöms i första hand utifrån projektets PICO (Kapitel 3) och huruvida de innehåller någon form av ekonomisk analys. Därefter granskas studiernas kvalitet och överförbarhet till svenska förhållanden och det sjukvårdssammanhang som frågeställningen gäller.

Kvaliteten på hälsoekonomiska utvärderingar beror främst på kvaliteten på data och vilka principer som använts för att beräkna kostnader och effekter. Den ekonomiska utvärderingen kan inte bli bättre än vad ingående data möjliggör. Hälsoekonomiska analyser har kritiserats för att beräkningsprinciperna varierar och för att det saknas en standard. För att bedöma kvaliteten har det därför utvecklats ett antal checklistor [8–11]. Den första och vanligast förekommande checklistan är den lista som utvecklades av Drummond och medarbetare [9]. Det finns emellertid andra liknande checklistor [8,10] och det har även utvecklats specifika checklistor för att bedöma kvaliteten på modeller [11]. SBU har, baserat på dessa checklistor och erfarenhet från tidigare arbete, utvecklat två egna mallar för granskning; en för empiriska studier och en för modellstudier (Bilaga 7–8). De har grunden gemensam men har anpassats för att bättre fånga de specifika frågor som gäller de olika typerna av studiedesign. De har även kompletterats med tre frågor om risk för jäv som även ingår i mallarna för granskning av medicinska studier. När det gäller bedömning av kvaliteten på data som används i modeller hänvisas till Cooper och medarbetare [12] som har gjort en hierarkisk kvalitetsbedömning av olika typer av data som kan användas i modeller.

När checklistorna används är det viktigt att komma ihåg att endast ett fåtal hälsoekonomiska analyser uppfyller checklistornas krav i sin helhet. Det innebär inte att studier som inte motsvarar alla krav skulle vara utan värde, men däremot att man bör vara medveten om bristerna vid tolkning av resultaten.

Egna analyser

Ibland visar det sig att den systematiska översikten inte kan besvara projektets ekonomiska frågeställningar. Detta gäller främst då tillgången på hälsoekonomiska studier inom aktuellt område är knapp och resultaten i tillgängliga empiriska utländska studier inte är relevanta i förhållande till svenska förutsättningar. Om det inom ramen för projektet går att få fram trovärdiga data avseende kostnader och effekter görs ibland egna analyser av kostnadseffektivitet. Oftast utgår analyserna från den kliniska evidens som fastslagits i projektet och kompletteras med beräkningar över vad de olika alternativen kostar. Beroende på komplexiteten i frågan och tillgången på data kan dessa analyser bli mer eller mindre omfattande. Om exempelvis evidens saknas för att effektskillnad finns mellan två behandlingar, och den ena är uppenbart mer kostsam behövs inga omfattande beräkningar för att påvisa att metoden inte är kostnadseffektiv. I andra fall kan omfattande analys av både kostnader och effekter behöva göras. I vissa fall kan det även bli aktuellt att göra egna modellanalyser. De inom ett projekt framtagna egna modellanalyserna görs vanligen med utgångspunkt från tillgängliga kliniska studier men anpassas till svenska förhållanden (t ex kostnadsdata, epidemiologi). Utformningen av beräkningarna bör följa samma krav som ställs för att publicera studier i vetenskapliga tidskrifter. Projektgruppens medicinska experter bör konsulteras för att bedöma om de i kalkylerna använda medicinska data är relevanta och korrekta. Oavsett form av modellanalys bör egna modellberäkningar bli föremål för utförlig känslighetsanalys och såväl intern som extern granskning.

Vad är en hälsoekonomisk utvärdering?

I hälsoekonomiska utvärderingar jämförs två eller flera alternativa behandlingsmetoder med avseende på såväl kostnader som effekter i syfte att klargöra vilken metod som är kostnadseffektiv i jämförelse med de andra alternativen [9]. Kostnadseffektivitet är alltså ett relativt begrepp. Ibland kan det mest relevanta alternativet emellertid vara ”ingen behandling”. Det är vanligt att skilja mellan fyra olika typer av hälsoekonomiska utvärderingar. Samtliga utvärderingar inkluderar kostnader men skiljer sig åt när det gäller beskrivning och värdering av effekter, se Tabell 11.1.

De tre förstnämnda typerna av utvärdering är egentligen varianter på samma metodik. Till skillnad från kostnadsintäktsanalysen mäter de inte effekterna i monetära termer. I *kostnadsminimeringsanalysen* förutsätts effekterna vara likvärdiga och alternativen jämförs därför endast med avseende på sina kostnader. I en *kostnadseffektanalys* används ett endimensionellt effektmått, såsom antal botade, antal besvärsfria dagar, antal överlevande eller antal vunna levnadsår (”life years saved”, LYS). *Kostnadsnyttoanalysen* innebär att kostnaderna relateras till ett nyttoindex, vanligen konstruerat som en sammanvägning av överlevnad och livskvalitet, t ex antalet vunna kvalitetsjusterade levnadsår (”quality

Tabell 11.1 Olika typer av hälsoekonomiska analysmetoder.

Typ av utvärdering	Effektmått	Kvot för analysen
Kostnadsminimeringsanalys *(Cost Minimisation Analysis, CMA)*	Inget effektmått då effekterna förutsätts vara helt lika	Presenterar bara kostnader
Kostnadseffektanalys *(Cost Effectiveness Analysis, CEA)*	Fysiska enheter, t ex levnadsår	Till exempel kostnad per vunnet levnadsår (LYS)
Kostnadsnyttoanalys *(Cost Utility Analysis, CUA)*	Nyttoindex, t ex kvalitetsjusterade levnadsår	Kostnad per vunnet kvalitets-justerat levnadsår (QALY)
Kostnadsintäktsanalys *(Cost Benefit Analysis, CBA)*	Pengar, t ex levnadsår värderade i monetära termer	Nyttan (i kronor) av vunnet levnadsår jämfört med kostnader för interventionen

adjusted life years", QALY). I praktiken görs ofta inte en uppdelning mellan kostnads-effektanalys och kostnadsnyttoanalys, utan kostnadsnyttoanalysen benämns som en kostnadseffektanalys med QALY som effektmått.

I en *kostnadsintäktsanalys* värderas även effekterna i pengar, vilket alltså direkt ger besked om den studerade behandlingens "lönsamhet". Denna typ av analys har länge ansetts svår eller omöjlig att tillämpa i sjukvårdssammanhang pga de praktiska svårigheterna att värdera effekterna i monetära termer. Senare års metodutveckling inom området, bl a genom olika metoder för att mäta betalningsvilja, har inneburit att kostnadsintäkts-analysen kommit till viss ökad användning, men fortfarande kvarstår metodologiska problem. Som en parallell till detta kan nämnas att både Vägverket och Banverket till-lämpar kostnadsintäktsanalys, inkluderande monetär värdering av liv (= värdet av ett statistiskt liv), för att beräkna den samhällsekonomiska lönsamheten av olika investe-ringsalternativ eller projekt inom infrastruktursektorn.

Olika typer av utvärderingar besvarar olika frågor. Valet av metod bestäms av aktuell frågeställning men även av tillgången på relevanta data. Om utvärderingen ska ligga till grund för val mellan två behandlingsmetoder (t ex alternativa läkemedel) med samma terapeutiska effekt och inga skillnader vad gäller biverkningar, så är det naturligt att nöja sig med en kostnadsminimeringsanalys. Handlar det om alternativa metoder som främst påverkar dödligheten kan det räcka att göra en kostnadseffektanalys med levnadsår som effektmått. Om det däremot rör sig om behandling av tillstånd som inte är direkt livsho-tande, t ex kroniska sjukdomar, är det nödvändigt att även beakta effekterna på livskvali-tet, vilket alltså pekar på kostnadsnyttoanalys som lämplig metod.

En ytterligare analysmetod som lanserats som ett komplement eller alternativ till de fyra ovan redovisade är så kallad kostnadskonsekvensanalys ("cost-consequences analysis",

CCA) [9], i vilken kostnader och effekter presenteras utan att adderas (se exempel för kostnadssidan i Tabell 11.2). En kostnadskonsekvensanalys överlåter åt beslutsfattaren att själv välja de för beslutssituationen mest relevanta uppgifterna och dra egna slutsatser därav.

Tabell 11.2 Kostnader i en kostnadskonsekvensanalys.

Kostnader	Nuvarande praxis	Ny intervention	Differens
Direkta kostnader • personalbemanning • läkemedel • avskrivningar • underhåll av utrustning • patienttid/anhörigtid • patienttransportkostnad • övriga Indirekta kostnader • produktionsförluster			

Vad menas med kostnadseffektivitet?

För att avgöra vilken av två metoder som är kostnadseffektiv behövs uppgifter om både kostnader och effekter. Om en ny metod har lägre kostnad och bättre effekt än jämförd använd metod så är den nya metoden, som det uttrycks, ”dominant” och valet av metod ter sig från hälsoekonomisk synpunkt enkelt. Oftast är emellertid effektivare metoder mer kostnadskrävande. I en beslutsmatris visas nedan (Tabell 11.3) de nio alternativ som kan uppkomma vid en jämförelse mellan metoder.

Tabell 11.3 En beslutsmatris för kostnadseffektivitet.

Ny metod jämförs med gammal	Sämre effekt	Lika effekt	Bättre effekt
Lägre kostnad	1. Läget oklart, *ev inkrementell analys*	4. Inför den nya metoden	7. Inför den nya metoden
Lika kostnad	2. Behåll den gamla metoden	5. Metoderna likvärdiga	8. Inför den nya metoden
Högre kostnad	3. Behåll den gamla metoden	6. Behåll den gamla metoden	9. Läget oklart, *gör inkrementell analys*

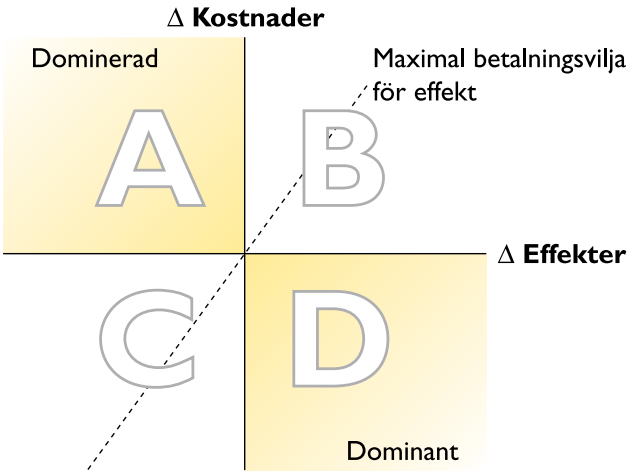
Vid alternativen 2, 3 och 6 är den gamla metoden kostnadseffektiv och behålls. Vid alternativen 4, 7 och 8 gäller motsatsen, dvs att den nya metoden är kostnadseffektiv. För alternativ 5 föreligger ingen skillnad, inget talar för ett behov av att byta till en nyare metod. Däremot behövs eventuellt för alternativ 1 men definitivt för alternativ 9 ytterligare analyser.

Resultatet från en hälsoekonomisk analys presenteras ofta som en inkrementell kostnads-effektkvot (”incremental cost-effectiveness ratio”, ICER), vilken är kvoten mellan kostnadsskillnad och effektskillnad:

ICER = (Kostnad A – Kostnad B) / (Effekt A – Effekt B)

där Kostnad A står för kostnaden som följer behandling A, Kostnad B står för kostnaden som följer behandling B, Effekt A står för effekten av behandling A och Effekt B står för effekten av behandling B. Alltså anger ICERn kostnaden för att uppnå ytterligare en enhet av effekt (t ex i form av vunna levnadsår) när man byter från den ena metoden till den andra. Den nya metoden är kostnadseffektiv om det i samhället finns en betalningsvilja för den ytterligare effekten som är högre än vad metoden kostar per effektenhet.

Resultatet kan även förklaras med hjälp av en ”kostnadseffektplan” (”cost-effectiveness plane”), dvs värdena för beräknade ICER kan förklaras i en tänkt matris med fyra kvadranter (Figur 11.1). Då kvadrant A respektive D ger självskrivna svar (enligt A dominerar den gamla metoden, enligt D dominerar den nya metoden) fokuseras intresset i allmänhet främst på kvadrant B och C, dvs då den nya metoden medför högre effekt men också högre kostnad eller lägre kostnad men sämre effekt jämfört med alternativet. Om man vet hur mycket samhället är berett att maximalt betala för en bättre behandling kan man rita in en gräns för vad som är kostnadseffektivt. Denna gräns går då igenom kvadranterna B och C och alla behandlingar som har en ICER till höger om denna linje uppfattas som kostnadseffektiva.



Figur 11.1 Kostnadseffektplan. A: högre kostnad/sämre effekt; B: högre kostnad/bättre effekt; C: lägre kostnad/sämre effekt; D: lägre kostnad/bättre effekt, jämfört med alternativet.

Val av perspektiv på analysen

Oftast eftersträvas att analysen ska använda ett samhällsperspektiv för att den ska visa de totala kostnaderna och effekterna för hela samhället och inte leda till suboptimering inom olika sektorer. Att ha ett samhällsperspektiv innebär att kostnader och effekter ska beaktas oberoende av var och när de uppkommer. Det kan emellertid ändå vara av intresse att beskriva hur kostnader och effekter fördelar sig på olika intressenter, såsom patient, landsting, kommun, staten m fl.

Kostnader

I en ekonomisk analys ingår både intäkter och kostnader uttrycka i monetära termer. Kostnader uppstår när resurser förbrukas för att ge en viss behandling. Om en behandling har en positiv effekt i form av livskvalitet och överlevnad kan detta också innebära framtida besparingar. Ett exempel är ett antirökningsprogram som medför en initialkostnad för själva programmet men som resulterar i framtida kostnadsbesparingar då samhället inte behöver lägga lika stora resurser på att behandla rökningsrelaterade sjukdomar.

Utifrån ett samhällsperspektiv bör samtliga relevanta kostnader förknippade med behandling och sjukdom identifieras, kvantifieras och värderas. Det inom hälsoekonomin relevanta kostnadsbegreppet är *alternativkostnaden*, dvs värdet av det som kan uppnås av resurserna i bästa alternativa användning. I praktiken är man dock oftast hänvisad till att använda marknadspriser eller kostnader härledda ur hälso- och sjukvårdens kostnadsredovisningar.

Kostnader relaterade till sjukdom och vård kan delas in i olika kategorier; direkta hälso- och sjukvårdskostnader, direkta övriga kostnader och indirekta kostnader [13]. Direkta kostnader är den resursförbrukning som uppstår som en direkt följd av vård och behandling, medan indirekta kostnader är de resurser som förloras indirekt pga sjukdom eller behandling, t ex nedsatt arbetsförmåga. Exempel på direkta kostnadsslag är personal, material, byggnader och läkemedel. I vissa fall kan kostnader i andra samhällssektorer än sjukvården, t ex primärkommunen, ha en stor betydelse för helhetsbilden. De viktigaste indirekta kostnaderna är produktionsbortfall pga sjuklighet eller för tidig död.

Underlag för att beräkna kostnader kan hämtas från svenska register eller statistikkällor. Socialstyrelsen har t ex hälsodataregister och statistikdatabaser som innehåller uppgifter om vårdtillfällen, antal operationer, vårddagar, medelvårdtider och läkemedelskonsumtion för olika åldersgrupper uppdelat på diagnoser, operationer eller DRG (diagnosrelaterade grupper). Dessa återfinns under www.socialstyrelsen.se/register respektive www.socialstyrelsen.se/statistik. Sveriges Kommuner och Landsting (SKL) har en kostnadsdatabas, KPP-databas, som innehåller uppgifter om kostnad per patient vid vissa sjukhus. Regionala priser och ersättningar publiceras av regionvårdsnämnder, t ex Södra Regionvårdsnämnden (www.srvn.org). Ytterligare en källa är de nationella kvalitetsregistren som ofta innehåller ganska specifika data om behandlingsinsatser och patientens status. Här måste man kontakta respektive register för att få mer detaljerad information. Mer information om KPP-databasen och en lista på kvalitetsregister som fått finansiellt stöd finns också på SKL:s webbplats (www.skl.se).

Att beräkna värdet av produktion

Kostnader för produktionsbortfall kan uppstå när en individ inte kan arbeta pga att han eller hon får en viss behandling. Ett vanligt scenario är emellertid att en behandling gör att någon som tidigare varit sjuk kan komma igång och arbeta igen, vilket alltså då resulterar i ökad produktion. Förutom när en individ är frånvarande från arbetet kan indirekta kostnader uppkomma om individen arbetar men till följd av sin sjukdom eller skada har lägre produktivitet än tidigare, vilket brukar beskrivas som ”sjuknärvaro” eller ”presenteeism”.

Det finns två olika metoder för att skatta värdet av produktion; humankapitalmetoden och friktionskostnadsmetoden [9]. Med humankapitalmetoden görs värderingen av produktionsbortfall vanligtvis under antagande att bortfallet kan värderas till marknadspris, dvs lön plus arbetsgivaravgifter. Ett problem med att skatta indirekta kostnader med humankapitalmetoden är att beräkningarna kan leda till överskattningar. Detta är en känd kontrovers i den hälsoekonomiska litteraturen, där flera holländska hälsoekonomer förespråkar friktionskostnadsmetoden (”friction cost methods”) [14,15]. Med ”friktion” menas den tid (med tillhörande kostnad) som går innan en tidigare arbetslös individ fullt ut kan ersätta en person. Metoden kan ses som ett pragmatiskt förhållningssätt till det faktum att produktionsförändringarna oftast är koncentrerade till kortvarig sjukfrånvaro eller till den första tiden av sjukfrånvaron, dvs tills en ersättare har övertagit arbetsuppgifterna.

För individer som inte är i arbetsför ålder (individer över 65 år) inkluderas vanligen inte produktionsförändringar i analysen. Detta har dock kritiserats då individer som inte längre är i arbetsför ålder ofta bidrar med informell produktion, vilket också borde värderas och inkluderas i den hälsoekonomiska analysen [16]. Att inkludera påverkan på produktion i analysen kan också anses stå i konflikt med människovärdesprincipen som säger att prioriteringar inom svensk hälso- och sjukvård ska göras ”oberoende av personliga egenskaper och funktioner i samhället” [16,17]. När behandlingars påverkan på produktion tas med i analysen endast för individer under 65 innebär det att åtgärder riktade till äldre skulle kunna komma att ges lägre prioritet. Det har därför rekommenderats att resultatet från hälsoekonomiska analyser presenteras både med och utan produktionskostnader [9,16].

Om löner används som underlag för att beräkna produktionskostnader är det viktigt att man inte gör detaljerade indelningar eftersom skillnader i lön mellan olika grupper kan bero på annat än värdet av vad som produceras. Ett exempel är skillnader i lön mellan män och kvinnor. Om produktionsbortfall för kvinnor respektive män beräknas olika baserat på officiell lönestatistik, kommer en behandling riktad till män att få en lägre

kostnad per effekt än om samma behandling ges till kvinnor, vilket alltså även skulle kunna leda till att behandlingar riktade till män ges högre prioritet. För att undvika detta kan produktionsförlusten beräknas som en genomsnittlig nationell lönekostnad istället för en genomsnittlig lönekostnad för varje kön.

Kvalitetsjusterade levnadsår (QALY)

Det rekommenderas ofta att den hälsoekonomiska analysen ska använda kvalitetsjusterade levnadsår (QALY) som effektmått [18–20]. QALYs mäter både överlevnad och livskvalitet, dvs både livslängd och hälsostatus liksom effekter av eventuella biverkningar, exempelvis 5 överlevnadsår med 0,7 livskvalitetsvikt innebär (5 x 0,7) = 3,5 QALY. Fördelen med QALYs är att de i princip kan användas för jämförelser mellan helt olika behandlingsområden. Detta kan emellertid vara problematiskt om det saknas tillräckligt säkra och generellt giltiga livskvalitetvikter, så kallade QALY-vikter.

QALY-vikter kan skattas med direkta och indirekta metoder. De direkta metoderna används för att skatta värdet av olika hälsotillstånd. Värdet skattas på en skala där 0 är lika med död och 1 är full hälsa. De indirekta metoderna består av ett frågeformulär som kan kopplas till en tariff som tagits fram med någon av de direkta metoderna.

De vanligaste direkta metoderna är ”standard gamble” (SG) [21], ”time trade-off” (TTO) [22] och ”visual analogue scale” (VAS) [23]. Alla kan användas såväl för att be patienter skatta sin egen livskvalitet som för att be allmänheten skatta hypotetiska tillstånd. SG och TTO är baserade på att individer får göra val mellan olika scenarion medan VAS bygger på att individer får värdera ett hälsotillstånd på en skala mellan bästa tänkbara tillstånd och sämsta tänkbara tillstånd.

De mest vanligt förekommande indirekta instrumenten är EQ-5D [24], SF-6D [25] och HUI-3 [26]. Frågeformulären de bygger på ser lite olika ut och tarifferna som används för att koppla svaren i formulären till QALY-vikter har tagits fram på olika sätt. De tariffer som oftast används idag bygger alla på värderingar gjorda av representanter från den allmänna befolkningen i antingen Storbritannien eller Kanada. Till exempel baseras den vanligaste tariffen för EQ-5D (som ofta används i Sverige) på värderingar av 3 395 britter som fått skatta värdet av olika kombinationer av svarsalternativ i EQ-5D formuläret med TTO-metoden [27].

De olika metoderna för att mäta QALY-vikter har inom ett flertal olika sjukdomsområden visats ge olika resultat [28–34], vilket i en ekonomisk utvärdering kan komma att påverka en behandlings kostnad per QALY. Detta kan bero på skillnader i vilka frågor som ingår i formulären, vilken direkt värderingsmetod som har använts, vilken statistisk

metod som använts för tariffen samt vilken population som fått skatta vikterna till tariffen. Då tariffer från olika länder har setts skilja sig åt [35], bör QALY-vikternas baseras på direkta mätningar i respektive land men det saknas ännu QALY-vikter baserat på svenska mätningar.

Värdet på en QALY – Hur vet man att en metod är kostnadseffektiv?

11

En behandling bedöms som kostnadseffektiv i förhållande till en annan om dess inkrementella kostnadseffektkvot är lägre än betalningsviljan för en QALY. Idealt skulle betalningsviljan för en QALY representera alternativkostnaden av att implementera behandlingen i fråga, vilket alltså är vad vi skulle förlora om vi omfördelade resurser från de behandlingar som redan ingår i hälso- och sjukvårdens budget till den nya behandlingen [18,36]. Om införandet av den nya metoden skulle innebära att vi måste utesluta en behandling som ger oss 1 QALY för 500 000 kronor måste alltså den nya behandlingen ge oss 1 QALY för mindre än 500 000 kronor för att ett införande ska kunna motiveras. Annars skulle det innebära att vi får ut mindre hälsa av ett införande än vad vi redan fick med dagens fördelning av resurser. För att ovanstående resonemang ska fungera kräver det att vi skulle veta kostnaden per QALY för allt som görs inom hälso- och sjukvården, vilket tyvärr är svårt om inte praktiskt omöjligt. Av dessa anledningar har det uppkommit andra ansatser för att skatta samhällets betalningsvilja för en QALY [36–39].

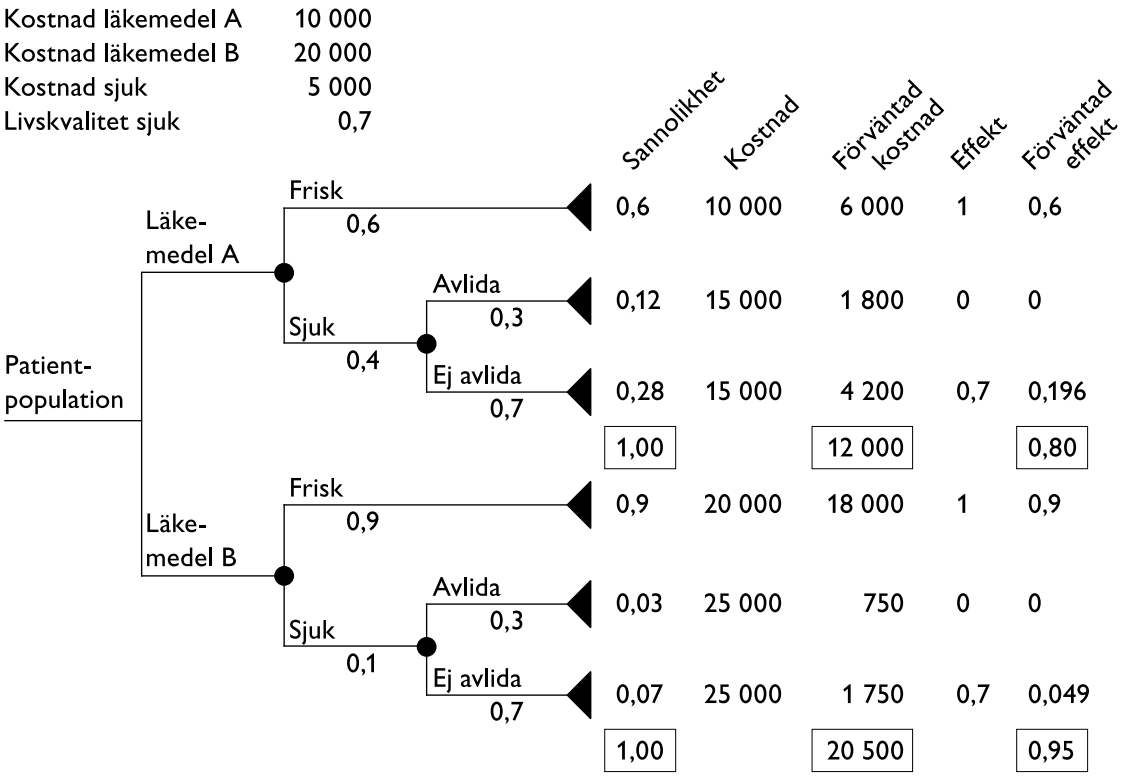
I England och Wales använder sig NICE av ett tröskelvärde på 20 000–30 000 brittiska pund för att bestämma huruvida en behandling anses som kostnadseffektiv [18]. I Sverige har det inte satts en exakt gräns för hur mycket en QALY får kosta för att behandlingen ska uppfattas som kostnadseffektiv. Dock har Socialstyrelsen i sina riktlinjer presenterat riktmärken för vad som uppfattas som en låg respektive hög kostnad per QALY. En låg kostnad per QALY definieras som under 100 000 kronor, en hög kostnad per QALY som över 500 000 kronor och en mycket hög kostnad per QALY som över 1 miljon kronor [40,41].

Modellanalyser

Ekonomiska utvärderingar inom sjukvården är i hög grad beroende av i vilken utsträckning kostnader och effekter av sjukvårdens behandlingar är kända. När så inte är fallet, t ex vid diskussion om att införa nya metoder för utredning eller behandling, kan bästa tillgängliga data sammanfattas i en så kallad modellanalys. En modell syftar till att belysa ett beslutsproblem utifrån bästa tillgängliga information, inte att ersätta empiriska studier. Det är främst i situationer när kostnader och effekter till följd av behandlingar uppstår under en längre tid än vad som har kunnat studeras i en studie som modeller tillämpas vid hälsoekonomisk utvärdering. Dessutom är det ofta aktuellt vid följande situationer [42]:

- Då relevanta kliniska utvärderingar saknas eller inte inkluderar data på kostnader och QALYs.
- För att extrapolera från intermediära utfallsmått.
- Då det av etiska skäl är omöjligt att genomföra kontrollerade kliniska prövningar.
- Då kostnaderna för att genomföra tillräckligt stora empiriska studier i vissa fall är orimligt höga i förhållande till det potentiella värdet av den ytterligare information som kan vinnas.
- Att kostnader som beräknats enligt kliniska prövningar inte är realistiska eller att de inte är relevanta för svensk sjukvård.

De vanligaste teknikerna vid modellanalyser inom hälsoekonomin är *beslutsträd* (Figur 11.2) och *Markov-modeller* (Figur 11.3) [42]. Principerna för dessa två metoder är i stora avseenden lika, men ett beslutsträd visar en sekvens av händelser under en bestämd tidsperiod. Denna teknik är lämplig vid utvärdering rörande sjukdomar av mer akut karaktär med ett händelseförlopp som är begränsat till en relativt kort tidsperiod.

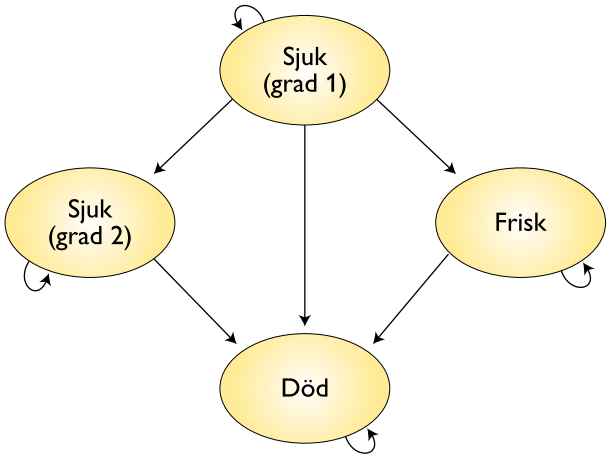


Figur 11.2 Beslutsträd. Exempel: jämförelse av två alternativa läkemedel A och B.

I Figur 11.2 jämförs två alternativa läkemedelsbehandlingar (A och B) med hjälp av ett beslutsträd, bestående av två beslutsgrenar som sedan förgrenar sig beroende på olika utfall av behandlingarna. Sannolikheten för olika utfall anges vid respektive gren. Samtliga grenar slutar i så kallade slutnoder (trianglar). I övre vänstra hörnet av figuren anges

ingångsvärden för aktuella parametrar. Till höger om trädet anges i första kolumnen sannolikheten för att hamna i respektive slutnod, givet det initiala valet av behandlingsstrategi. I övriga kolumner anges på motsvarande sätt kostnad, förväntad kostnad, effekt och förväntad effekt. Inramade värden i tredje och femte kolumnerna anger förväntad kostnad och förväntad effekt av de två alternativen A och B. Den inkrementella kostnadseffektkvoten (ICER), dvs merkostnaden per effektenhet om man väljer läkemedel B istället för A, blir (20 500–12 000)/(0,95–0,80) = 56 667 kronor.

Markov-modeller är uppbyggda kring ömsesidigt uteslutande hälsotillstånd, se Figur 11.3. Man definierar övergångsannolikheter för förflyttningar mellan de olika hälsotillstånden, vilka också är förenade med vissa kostnader och hälsoutfall. Markov-modellen är mer praktiskt användbar för analys av beslutsproblem som avser lång tid, t ex behandling av kronisk sjukdom.



Figur 11.3 Markov-modell.

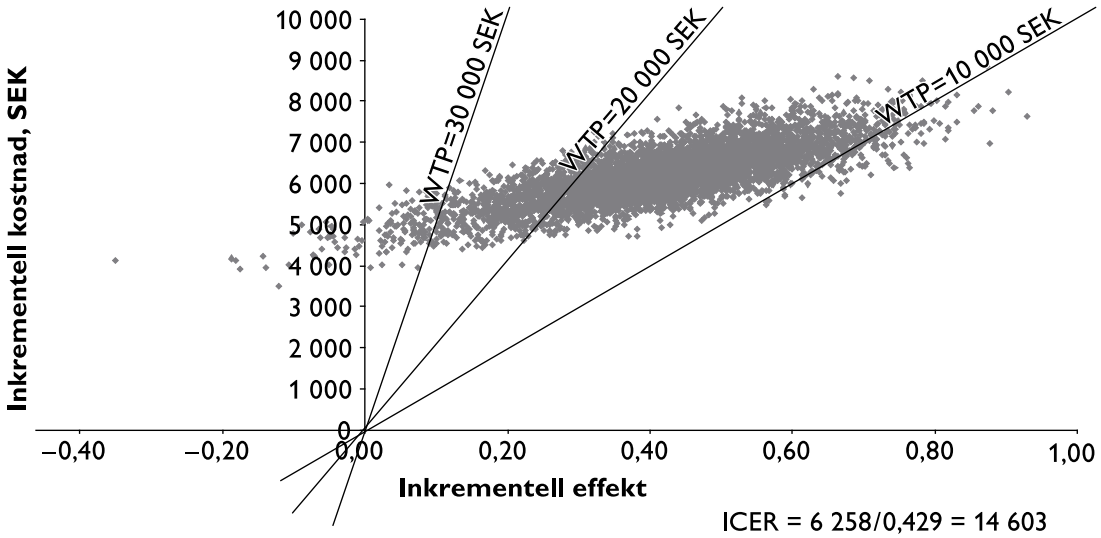
Förutom beslutsträd och Markov-modeller har det på senare tid även blivit vanligare att använda sig av händelsestyrda modeller (”discrete event simulation”, DES) [43]. Istället för att utgå från olika hälsotillstånd som i Markov-modellerna, bygger dessa modeller på olika händelser (”events”) som inträffar vid specifika tidpunkter. Det kan vara händelser som t ex att en patient insjuknar, ett läkarbesök eller att behandling påbörjas. Flera olika händelser kan ske samtidigt och var och en av dessa händelser kan i sin tur få konsekvenser i form av t ex kostnader, livskvalitetsförändringar och/eller förändrad risk för framtida händelser. DES-modellerna har dock kritiserats för att de kräver mer detaljerade data, vilka ofta inte är publicerade och kan vara svåra att få tag på. En annan kritik mot DES-modellerna är att det har ansetts svårt och tidskrävande att utföra så kallade probabilistiska känslighetsanalyser i dessa modeller [44] men den uppfattningen delas inte av alla [43].

Känslighetsanalyser

Vid hälsoekonomiska utvärderingar är det ofta viktigt att göra känslighetsanalyser [13]. Att göra en känslighetsanalys innebär att man varierar en eller flera variabler i analysen för att undersöka vad som händer med analysens resultat. Det kan t ex gälla att öka interventionskostnaden om man misstänker att den kan vara högre än den skattning man använt i sin grundanalys. Det kan också handla om att man är osäker på olika

antaganden som gjorts i en modell, t ex hur stor andel av patienterna som behöver göra om en undersökning eller hur ofta en patient behöver komma tillbaka för kontrollbesök. Då kan man testa ett flertal olika scenarion och se om resultatet ändras eller inte. En annan ansats är att analysera vid vilket värde (tröskelvärde) på variabeln som resultatet om kostnadseffektivitet ändras, dvs vid vilket variabelvärde som en metod som var kostnadseffektiv inte längre är kostnadseffektiv eller tvärtom.

I modeller brukar probabilistisk känslighetsanalys (”probabilistic sensitivity analysis”) tillämpas [42]. Det förkortas ofta ”PSA” och innebär att osäkerheten kring modellens variabler inkluderas i analysen. Varje variabel får då en statistisk fördelning (t ex normal-, beta- eller gammafördelning) utifrån den osäkerhet som omger den specifika variabeln (t ex baserat på uppgifter om standardavvikelse). Därefter körs modellen flera gånger (ofta mellan 1 000 och 10 000 gånger) varvid olika tänkbara variabelvärden kombineras för beräkning av en förväntad kostnad per effekt. I varje körning dras ett värde från varje variabelfördelning och ett resultat beräknas. I Figur 11.4 illustreras resultatet av en modell som körts 5 000 gånger. Linjerna i figuren anger olika nivåer för maximal betalningsvilja för en effektenhet. Förutom medelvärdet av alla skattningar presenteras i en PSA sannolikheten för att metoden är kostnadseffektiv. Den beräknas utifrån hur många procent av skattningarna som hamnar till höger om den linje som representerar betalningsviljan för en effekt. Till exempel visar figuren att cirka 90 procent av skattningarna hamnar till höger om linjen som representerar en betalningsvilja på 30 000 kronor per effektenhet, alltså är sannolikheten för att metoden är kostnadseffektiv cirka 90 procent om vi är beredda att betala 30 000 kronor för att vinna ytterligare en effektenhet. En liknande osäkerhetsanalys kan även göras på data från empiriska studier med hjälp av att bootstrappa kostnader och effekter [9].



SEK = Svenska kronor; WTP = ”Willingness to pay” (betalningsvilja)

Figur 11.4 Kostnadseffektplan med probabilistisk känslighetsanalys.

Kostnadseffektivitet vid diagnostik

Det finns få publicerade studier av kostnadseffektivitet vid diagnostik. En orsak till detta kan antas vara att det inte finns någon intressent som är villig att betala för studier inom diagnostik på samma sätt som läkemedelsföretag har intresse av att finansiera behandlingsstudier. Det finns emellertid anledning att uppmärksamma vissa hälsoekonomiska aspekter som är specifika just för diagnostik. Ett diagnostiskt test med låg specificitet kommer att leda till många falskt positiva fall vilket kan medföra att patienter oroas i onödan, att de blir föremål för ytterligare utredningar innan ett friande test erhålls och i värsta fall att de behandlas i onödan. Ett test med låg sensitivitet kommer att medföra många falskt negativa fall vilket kan medföra onödigt lidande för patienter, ökade behandlingskostnader när sjukdomen väl blir fastställd, och i värsta fall att patienter hinner avlida i brist på insatt adekvat behandling.

Mot bakgrund av ovanstående scenario är en given fråga vilket utfallsmått som ska användas vid beräkning av kostnadseffektivitet vid hälsoekonomisk utvärdering av diagnostik? En uppfattning är att kostnaden bör relateras till antalet korrekt diagnostiserade fall, dvs summan av sant positiva fall och sant negativa fall (motsvarande ”accuracy”). Vidare är en väsentlig fråga vilka kostnader som ska inkluderas vid beräkning av kostnadseffektivitet? Att kostnader för den diagnostiska undersökningen ska inkluderas är självklart, men ska även kostnader för falskt positiva fall (fortsatta utredningar; patienters oro) respektive falskt negativa fall (patienters lidande; ökade behandlingskostnader vid försenad diagnos; eventuellt förtida dödsfall) inräknas? Det finns ingen generell regel för hur man ska gå till väga vid beräkning av kostnadseffektivitet vid diagnostik, men givet ett samhällsperspektiv ska samtliga konsekvenser som är en följd av diagnostiken inkluderas.

En av de få studier som har uppmärksammat de specifika frågorna vid beräkning av kostnadseffektivitet för diagnostik är gjord av Laking och medarbetare år 2006 [45] och har olika ”cut-off”-nivåer för specificitet och sensitivitet som utgångspunkt, vilka kan tillämpas vid beräkning av ROC-kurvor.

Analys av budgetpåverkan (”budget impact analysis”)

För att underlätta för dem som ska finansiera införandet av en viss metod kan kostnadseffektanalyserna kompletteras med en budgetpåverkananalys (”budget impact analysis”). Analysen skiljer sig från de andra analysformerna på så sätt att den är till för att utvärdera hur en viss eller flera budgetar påverkas av införandet av en metod och vilka konsekvenserna förväntas bli för olika aktörer. Den har alltså inte i syfte att utvärdera om det finns en rimlig relation mellan metodens kostnader och effekter och går därför inte att använda för att optimera samhällets resurser. ISPOR Task Force har nyligen publicerat riktlinjer för budgetpåverkananalyser [46].

Hälsoekonomi och evidens

Hälsoekonomiska utvärderingar är teoretiskt baserade på ämnet nationalekonomi. Det innebär i sin tur att det bygger på teorier om människors beteenden och värderingar. Utfallen av utvärderingarna är därför inte med automatik en absolut sanning som är oberoende av tid och rum. Tanken med hälsoekonomiska utvärderingar är att de ska användas som stöd vid beslutsfattande och inte som evidens. Av den anledningen görs ofta andra analyser och statistiska test än de som genomförs för att fastställa en medicinsk åtgärds kliniska effekt.

Modeller som bygger på ett flertal olika källor och antaganden ska inte tolkas som evidens utan som en prognos om en metods balans mellan kostnader och effekter. Däremot är det viktigt att de effektmått som modellen bygger på är statistiskt säkerställda. Hälsoekonomiska utfallsmått i randomiserade kontrollerade studier kan evidensgraderas precis som de medicinska utfallsmåtten men det är viktigt att evidensgraderingen görs på de enstaka utfallsmåtten och inte den sammanvägda ICERn som består av en sammanslagning av ett flertal olika utfallsmått [8].

Avslutningsvis är det värt att påminna om att hälsoekonomiska analyser är viktiga för att kunna styra begränsade resurser inom hälso- och sjukvården på ett effektivt sätt. Om resurser används till åtgärder som inte är kostnadseffektiva, betyder det i förlängningen att resurserna tas ifrån en annan åtgärd som ger mer hälsa per krona och att vi då alltså inte använder resurserna på ett optimalt sätt.

Referenser

1. International Network of Agencies for Health Technology Assessment. HTA Resources. [citerad 25:e september 2012]; Tillgänglig från: http://www.inahta.net

2. Stevens A, Milne R, Burls A, Health technology assessment: history and demand. J Public Health Med, 2003;25:98-101.

3. Newdick C. Who should we treat? Rights, rationing, and resources in the NHS. Vol. 2nd edition. New York: Oxford University Press; 2005.

4. Rice DP. Estimating the cost of illness. Am J Public Health Nations Health, 1967;57:424-40.

5. Hodgson TA, Meiners MR. Cost-of-illness methodology: a guide to current practices and procedures. Milbank Mem Fund Q Health Soc, 1982;60:429-62.

6. Drummond M. Cost-of-illness studies: a major headache? Pharmacoeconomics, 1992;2:1-4.

7. Byford S, Torgerson DJ, Raftery J. Economic note: cost of illness studies. BMJ, 2000;320:1335.

8. Brunetti M, Ruiz F, Lord J, et al. Chapter 10: Grading economic evidence. In: Shemilt I, Mugford M, Vale L, et al, editors. Evidence-based decisions and

economics: health care, social welfare, education and criminal justice. Oxford: Wiley-Blackwell; 2010.

9. Drummond MF, Sculpher MJ, Torrance GW, et al. Methods for the economic evaluation of health care programmes. Oxford: Oxford University Press; 2005.

10. Evers S, Goossens M, de Vet H, van Tulder M, Ament A. Criteria list for assessment of methodological quality of economic evaluations: Consensus on health economic criteria. Int J Technol Assess Health Care 2005;21:240-5.

11. Philips Z, Ginnelly L, Sculpher M, Claxton K, Golder S, Riemsma R, et al. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. Health Technol Assess, 2004;8:iii-iv, ix-xi, 1-158.

12. Cooper N, Coyle D, Abrams K, Mugford M, Sutton A. Use of evidence in decision models: an appraisal of health technology assessments in the UK since 1997. J Health Serv Res Policy 2005;10:245-50.

13. Gold MR, Siegel JE, Russell LB, Weinstein MC, editors. Cost-effectiveness in health and medicine. New York: Oxford University Press; 1996.

14. Johannesson M, Karlsson G. The friction cost method: a comment. J Health Econ 1997;16:249-55; discussion 257-9.

15. Koopmanschap MA, Rutten FF, van Ineveld BM, van Roijen L. The friction cost method for measuring indirect costs of disease. J Health Econ 1995;14:171-89.

16. Sahlén K-G, Löfgren C, Lindholm L, Är det lönsamt med prevention efter 65? Ålderns betydelse i hälsoekonomiska utvärderingar. Stockholm: Statens folkhälsoinstitut; 2006. R 2006:19. ISBN 91-7257-447-X.

17. Sculpher M. The role and estimation of productivity costs in economic evaluation. In: Economic evaluation in health care: Merging theory with practice. Drummond M, McGuire A, editors. Oxford: Oxford University Press; 2001.

18. National Institute for Health and Clinical Excellence. Guide to the methods of technology appraisal. London: National Institute for Health and Clinical Excellence (NICE); 2008.

19. International Society for Pharmacoeconomics and Outcomes Research. Pharmacoeconomic guidelines around the world. 2008 [citerad 25:e september 2012]; Tillgänglig från: http://www.ispor.org/PEguidelines/index.asp

20. Läkemedelsförmånsverket (LFN). Läkemedelsförmånsnämndens allmänna råd om ekonomiska utvärderingar. 2003. LFNAR 2003:2.

21. von Neumann J, Morgenstern O. Theory of games and economic behavior. Princeton, NJ: Princeton University Press; 1944.

22. Torrance GW, Thomas WH, Sackett DL. A utility maximization model for evaluation of health care programs. Health Serv Res 1972;7:118-33.

23. Patrick DL, Bush JW, Chen MM. Methods for measuring levels of well-being for a health status index. Health Serv Res 1973;8:228-45.

24. The EuroQol Group. EuroQol – a new facility for the measurement of health-related quality of life. The EuroQol Group. Health Policy 1990;16:199-208.

25. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002;21:271-92.

26. Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. Med Care 2002;40:113-28.

27. Dolan P, Gudex C, Kind P, Williams A. A social tariff for EuroQol: Results from a UK general population survey. York: Centre for Health Economics, University of York; 1995.

28. McDonough CM, Tosteson AN. Measuring preferences for cost-utility analysis: how choice of method may influence decision-making. Pharmacoeconomics 2007;25:93-106.

29. Seymour J, McNamee P, Scott A, Tinelli M. Shedding new light onto the ceiling and floor? A quantile regression approach to compare EQ-5D and SF-6D responses. Health Econ 2010;19:683-96.

30. Kopec JA, Willison KD. A comparative review of four preference-weighted measures of health-related quality of life. J Clin Epidemiol 2003;56:317-25.

31. Heintz E, Wiréhn AB, Peebo BB, Rosenqvist U, Levin LÅ. QALY weights for diabetic retinopathy – a comparison of health state valuations with HUI-3, EQ-5D, EQ-VAS, and TTO. Value Health 2012;15:475-84.

32. Bleichrodt H, Johannesson M, Standard gamble, time trade-off and rating scale: experimental results on the ranking properties of QALYs. J Health Econ 1997;16:155-75.

33. Puhan MA, Schünemann HJ, Wong E, Griffith L, Guyatt GH. The standard gamble showed better construct validity than the time trade-off. J Clin Epidemiol 2007;60:1029-33.

34. Stiggelbout AM, Kiebert GM, Kievit J, Leer JW, Stoter G, de Haes JC. Utility assessment in cancer patients: adjustment of time tradeoff scores for the utility of life years and comparison with standard gamble scores. Med Decis Making 1994;14:82-90.

35. Karlsson JA, Nilsson JÅ, Neovius M, Kristensen LE, Gülfe A, Saxne T, Geborek P. National EQ-5D tariffs and quality-adjusted life-year estimation: comparison of UK, US and Danish utilities in south Swedish rheumatoid arthritis patients. Ann Rheum Dis 2011;70:2163-6.

36. McCabe C, Claxton K, Culyer AJ. The NICE cost-effectiveness threshold: what it is and what that means. Pharmacoeconomics, 2008;26:733-44.

37. Rawlins MD, Culyer AJ. National Institute for Clinical Excellence and its value judgments. BMJ 2004;329:224-7.

38. Devlin N, Parkin D. Does NICE have a cost-effectiveness threshold and what other factors influence its decisions? A binary choice analysis. Health Econ 2004;13:437-52.

39. Smith RD, Richardson J. Can we estimate the 'social' value of a QALY? Four core issues to resolve. Health Policy 2005;74:77-84.

40. Socialstyrelsen. Bilaga 4, Metod. Nationella riktlinjer för diabetesvården 2010 – Stöd för styrning och ledning. Stockholm: Socialstyrelsen; 2010. ISBN 978-91-86301-88-0.

41. Socialstyrelsen. Metodbilaga, Metod för Socialstyrelsens arbete med nationella riktlinjer. Nationella riktlinjer för psykosociala insatser vid schizofreni eller schizofreniliknande tillstånd 2011 – stöd för styrning och ledning. Stockholm: Socialstyrelsen; 2011. ISBN 978-91-86585-77-8.

42. Briggs A, Claxton K, Sculpher M. Decision modelling for health economic evaluation. New York: Oxford University Press; 2006.

43. Caro JJ, Möller J, Getsios D. Discrete event simulation: the preferred technique for health economic evaluations? Value Health 2010;13:1056-60.

44. Claxton K, Sculpher M, McCabe C, Briggs A, Akehurst R, Buxton M, et al. Probabilistic sensitivity analysis for NICE technology assessment: not an optional extra. Health Econ 2005;14:339-47.

45. Laking G, Lord J, Fischer A. The economics of diagnosis. Health Econ 2006; 15:1109-20.

46. Mauskopf JA, Sullivan SD, Annemans L, Caro J, Mullins CD, Nuijten M, et al. Principles of good practice for budget impact analysis: report of the ISPOR Task Force on good research practices – budget impact analysis. Value Health 2007;10:336-47.

12. Etiska och sociala aspekter

VERSION 2011:1.1

Bakgrund

Etiska och sociala aspekter bör beaktas i alla led av utvärderingsarbetet, alltifrån valet av utvärderingsobjekt och sedan fortlöpande genom hela utvärderingsprocessen. Tyngdpunkten ligger dock på att belysa etiska och sociala aspekter på tillämpningen av en viss medicinsk metod och det är främst detta som tas upp här [1,2].

Etiska aspekter

Vad menas med etiska aspekter?

Etik (från grekiskans *ethos; sed*) är liktydigt med moralfilosofi, dvs den del av den filosofiska vetenskapen som försöker besvara frågor som ”vad är det goda”, ”vad är det rätta”, ”hur bör man bete sig”. Etiska aspekter i sjukvårdssammanhang rör i första hand vad som gagnar eller skadar den enskilda patienten. Vidare aktualiseras frågor rörande respekten för patientens autonomi och integritet, samt rättvisa när det gäller vem som ska erbjudas vårdinsatser av olika slag. Detta kan sammanfattas i följande etiska principer för hälso- och sjukvården [3,4]:

1. Principen att göra gott, vilket innebär att man bör sträva efter att alltid försöka hjälpa och tillgodose en patients medicinska behov.

2. Principen att inte skada, som utgår ifrån att strävan att göra gott också ibland innebär ett förutsett risktagande för en patient och att man därför bör sträva efter att minimera risktagandet och åtminstone inte medvetet utsätta en patient för skada eller risk att skadas.

3. Autonomiprincipen säger att man bör visa respekt för en patients rätt att ha ett reellt inflytande på handlingar och beslut som berör patienten själv, samt för patientens värderingar, önskningar och åsikter. Det senare betraktas ibland som en särskild princip och benämns ”integritetsprincipen”.

4. Rättviseprincipen säger att man bör behandla och bemöta alla patienter med samma behov lika, dvs att det är patientens medicinska behov som ska avgöra hur man handlar, inte patientens position i samhället eller andra icke-medicinska aspekter.

Genomgående är det alltså patientperspektivet som står i fokus. Allt som görs för patienterna bör följa dessa principer. I praktiken uppkommer dock ofta konflikter mellan principerna. Exempelvis kan en strävan efter ökad autonomi för patienterna förväntas leda till en mer efterfrågeanpassad vård, men innebär samtidigt risk för en mindre rättvis fördelning av vårdkonsumtionen. Ett annat exempel på tänkbar konflikt mellan de olika principerna är om patienten inte är kapabel att förstå sitt eget bästa när det gäller val av behandling (autonomi kontra göra gott) (Exempel 12.1).

12

Exempel 12.1 Etiska problem kring metoder som ska utvärderas.

Många äldre som vistas på institution har dålig munhälsa. Personalen kan då uppleva en konflikt mellan å ena sidan principen om att göra gott (hjälpa patienterna med deras munhygien) och autonomiprincipen (avstå från att tvinga eller övertala motvilliga patienter att ta emot hjälp).

Det finns god evidens för att rådgivning om fysisk aktivitet till överviktiga patienter leder till ökad fysisk aktivitet vilket i sin tur kan ha positiva hälsoeffekter. Samtidigt kan ansvarig personal uppleva att de genom att – i enlighet med principen om att göra gott – aktivt ge råd om denna typ av livsstilsförändringar riskerar att inkräkta på patientens autonomi och integritet.

De etiska principerna relaterar emellertid inte enbart till den enskilda patientens rättigheter. Även andra aktörer berörs, såsom andra patienter eller patientgrupper, sjukvårdspersonalen, sjukvården som helhet eller samhället. Eftersom dessa aktörer kan ha olika intressen finns det risk för intressekonflikter.

Hur ska etiska aspekter beaktas och redovisas?

Etiska aspekter kan bli aktuella i olika led av SBU:s projektarbete. I projektplanen anges hur de etiska aspekterna ska hanteras i projektet och en arbetsmodell utarbetas som förslagsvis omfattar följande steg:

- Inledande analys
- Litteraturgranskning
- Etisk analys av utvärderingsområdet
- Diskussion med berörda grupper
- Slutsatser
- Förväntade praxisförändringar, implementering.

Inledande analys

I varje projekt ska i samband med formulering av frågorna övervägas vilka etiska aspekter som kan och bör belysas inom projektet. Som ett första led i arbetet är det alltså angeläget att – förslagsvis vid något av de första projektgruppsmötena – identifiera och beskriva tänkbara etiska frågor och problem kring de metoder som ska utvärderas. Detta kan ske med utgångspunkt från de fyra ovan beskrivna etiska principerna och med ledning av ett urval av frågorna i Hofmanns artikel [2]. SBU tillämpar i sitt arbete en checklista som bygger på några av Hofmanns frågor som man bedömt vara särskilt viktiga. De etiska frågornas tyngd kan variera mellan olika projekt beroende på i vad mån ämnet är etiskt kontroversiellt. I vissa fall, då det finns en tydlig etisk problematik (t ex genterapi), krävs som regel hjälp av en etiker redan från början och det är då lämpligt att en etiker ingår i projektgruppen.

Om möjligt bör analysen även belysa dagens praxis utifrån ett etiskt perspektiv. Kartläggning av praxis ska i princip ingå i alla utvärderingsprojekt. Detta kommer erfarenhetsmässigt ofta att visa på betydande variationer i praxis. Det kan då vara motiverat att formulera och försöka besvara frågor av typen:

- Är de metoder som tillämpas idag etiskt försvarbara med hänsyn till avvägningen mellan nytta, risk och kostnader? (Att använda ineffektiva metoder är slöseri med begränsade resurser och därmed oetiskt.)
- Är förekommande praxisskillnader förenliga med ”vård på lika villkor”? (Exempel 12.2)

Svaren på denna typ av frågor kan utgöra en viktig utgångspunkt för den etiska analysen av utvärderingens slutsatser och de konsekvenser som följer av dessa.

Exempel 12.2 Etiska problem vid påvisade praxisskillnader.

Inom glaukomvården uppvisar behandlingspraxis stora variationer, såväl vad gäller laserbehandling och invasiv kirurgi som läkemedelsbehandling. Frekvensen laserbehandlingar varierade år 2006 från 14 till 191 per 100 000 invånare mellan landstingen. För glaukomoperationer var motsvarande variationsvidd 2–74. Är detta vård på lika villkor? Är det förenligt med rättviseprincipen?

Litteraturgranskning

Sökning och granskning av vetenskaplig litteratur

Sökning av relevant litteratur med utgångspunkt från de etiska frågor som identifierats vid den inledande analysen görs i samarbete med SBU:s informationsspecialist. Lämpliga databaser, förutom de medicinska, kan vara IBSS och PsycINFO. Identifierade studier granskas och bedöms med avseende på studiekvalitet och relevans. För granskning av studier hänvisas till Kapitel 6–8.

Etiska aspekter på litteratur som handlar om effekter, biverkningar och kostnadseffektivitet

Under arbetet med granskningen av de vetenskapliga medicinska och hälsoekonomiska studierna bör projektgruppen ställa sig frågan om den bakomliggande forskningen bedrivits på ett etiskt acceptabelt sätt, t ex enligt den internationellt erkända Helsingforsdeklarationen [5]. En checklista för bedömning av forskningsetiska aspekter har föreslagits av Weingarten och medarbetare [6]. Där framhålls att man bör notera om de patienter som deltagit i en studie fått adekvat information och givit sitt samtycke, om etisk kommitté granskat och godkänt studien samt hur forskningen finansierats. Eventuell koppling till kommersiella och andra särintressen bör också noteras.

I princip bör även resultat från studier av detta slag användas, förutsatt att de håller tillräcklig kvalitet och resultaten bedöms relevanta och värdefulla.

Projektgrupperna ska formulera och diskutera de forskningsetiska problemen. Diskussionen kan eventuellt också sammanfattas i kapitlet om framtida forskning i respektive rapport (Exempel 12.3).

Exempel 12.3 Forskningsetiskt problem.

I studier av sockerersättningsmedel för att undvika karies fick försökspersonerna, som var barn i 10–14-årsåldern, tugga tuggummi med sockerersättningsmedel flera gånger om dagen. På detta sätt vande sig barnen vid daglig konsumtion av sött tuggummi. Studierna genomfördes i utvecklingsländer där produkter som innehåller sockerersättningsmedel är dyra och man kan befara att barnen istället fortsatte med sackarosinnehållande tuggummi efter studiens slut.

Etisk analys av utvärderingsområdet

När de etiska frågorna, eventuellt med hjälp av praxiskartläggning och litteratursökning, har identifierats av projektgruppen, kan en etiker leda den vidare diskussionen i gruppen om bästa sättet att belysa dessa frågor (Exempel 12.4).

Exempel 12.4 Etiska konsekvenser av utvärderingsresultat.

Alla överväganden om fosterdiagnostik rymmer komplexa etiska ställningstaganden. Utgångspunkten för SBU:s projekt var inte *om* det ska finnas tidig fosterdiagnostik eller ej, utan *hur* denna lämpligen ska vara utformad och vilka metoder som bör användas givet att den ska finnas. De etiska frågorna kom därför främst att handla om information, kunskap, beslutsfattande och psykologiska aspekter. Tidig fosterdiagnostik måste för att vara etiskt försvarbar bedrivas på ett sådant sätt att den av kvinnan/paret upplevs som frivillig, som ett erbjudande de kan välja att låta sig informeras om eller ej, delta i eller avstå ifrån. Nödvändiga förutsättningar är att informationen är tillräckligt omfattande, allsidig och neutral samt förmedlas av för ändamålet kvalificerad personal som inte otillbörligt påverkar kvinnans/parets ställningstaganden.

Etiska frågeställningar som generellt sett kan vara angelägna att analysera är [7]:

- Uppvägs risken för biverkningar/skador av den nytta behandlingen förväntas medföra?
- Ges patienten utförlig och tydlig information, och möjlighet att påverka vårdens utformning?
- Kan ianspråktagna resurser ge mer och/eller rättvisare fördelad nytta i annan användning? Ger den hälsoekonomiska analysen upphov till etiska frågor?

Syftet med analysen bör vara att klargöra vilka etiska konsekvenser de olika alternativen medför för berörda parter. De etiska konsekvenserna värderas utifrån de olika etiska principerna, vilket t ex kan ske med hjälp av Hermeréns så kallade aktörsmodell [8]. Enligt denna omfattar den etiska analysen följande steg: Problem – Faktabakgrund – Aktörer och berörda – Alternativ – Konsekvenser. Se Exempel 12.5 för ett exempel på etisk konsekvensanalys baserad på denna modell.

Exempel 12.5 Etisk analys avseende ”Rörbehandling av barn med sekretorisk mediaotit (SOM)”, baserad på Hermeréns aktörsmodell [8].

Problem, faktabakgrund: De flesta barn har haft minst en akut mediaotit och minst en episod med vätska i mellanörat. Sekretorisk mellanöreinflammation är i princip självläkande och det kan därför ifrågasättas om t ex rörbehandling är nödvändig och etiskt försvarbar. Skälet till att man överväger att behandla är främst att barnet under den tid inflammationen pågår kan ha nedsatt hörsel och sänkt livskvalitet.

Aktörer och berörda: Barnet, föräldrarna, samhället.

Alternativ: Alternativet till rörbehandling vid SOM är aktiv expektans.

Konsekvenser: Se Tabell 12.5.1.

Tabell 12.5.1 Etisk analys.

Aktörer	Etiska principer			
	Göra gott	**Inte skada**	**Autonomi**	**Rättvisa**
Barnet	Ökad livskvalitet Bättre hörsel	Obehag och risker med operationen	Barnets föräldrar övertar autonomin	
Föräldrarna	Bättre livskvalitet			
Samhället	Eventuellt minskat produktionsbortfall			Mindre resurser till behandling av mer allvarliga tillstånd

Analysen tydliggör att rörbehandling vid SOM ger upphov till konflikter mellan de olika etiska principerna. De positiva effekterna av behandlingen (högre livskvalitet och bättre hörsel för barnet, högre livskvalitet för föräldrarna) är kortsiktiga eftersom tillståndet även utan behandling läker ut så småningom. Barnet utsätts för obehag av operationen och kan själv endast i begränsad utsträckning påverka beslutet. Behandlingen tar i anspråk resurser, som skulle kunna användas för högre prioriterade behov inom sjukvården.

SBU:s projektgrupp kom fram till att rörbehandling är etiskt motiverad för barn med påtagliga besvär av SOM – med hänsyn till fördelarna för barnet och föräldrarna samt den låga skaderisken.

Diskussion med berörda grupper

I vissa fall kan det vara lämpligt att de grupper som kommer att vara berörda av rapporten och de etiska implikationerna av resultaten, t ex patienter, anhöriga och olika personalgrupper får ta del av och lämna synpunkter på den etiska analysen av utvärderingsområdet innan rapporten publiceras [7]. SBU:s lekmannaråd kan eventuellt involveras i detta arbete. Det kan också vara lämpligt att en professionell etiker ingår i gruppen av externa granskare av rapportmanus.

Slutsatser

Redovisningen av etiska aspekter bör komma till uttryck i slutsatserna och grundas på analys av frågor rörande risk–nytta, information/delaktighet respektive resursfördelning.

Förväntade praxisförändringar, implementering

De etiska frågorna och analysen av dessa bör också ses som ett resultat av utvärderingen och kan vara en del av de viktigaste budskapen i en så kallad kommunikationsplan. I den mån en konsekvensanalys genomförs inom ramen för projektet bör konsekvenserna analyseras även ur etiskt perspektiv.

Sociala aspekter

Vad menas med sociala aspekter?

Sociala aspekter kan avse såväl orsaker till sjukdom som konsekvenser av sjukdom och av användning av olika medicinska metoder. En central fråga är den om vård på lika villkor, dvs om vården, vid samma behov av insatser, ges oberoende av kön, ålder, ekonomisk och social situation eller bostadsort.

Sociala *orsaker* till sjukdom omfattar både strukturella och individuella faktorer som har betydelse för att en person blir sjuk, och kan vara avgörande för samhällets, sjukvårdens eller den enskildes möjligheter att påverka sjukdomsutvecklingen och förebygga komplikationer. Incidens och dödlighet i olika sjukdomar varierar nästan undantagslöst med sociala skillnader. Bristande tillgång till adekvat prevention för hela eller delar av befolkningen kräver andra insatser än om problemet har att göra med den enskilde individens beteende. Samhällsinriktade preventiva insatser kan ibland vara ett alternativ till sjukvårdsinsatser av det slag som de systematiska litteraturöversikterna oftast har fokuserat på (Exempel 12.6).

Exempel 12.6 Sociala orsaker till sjukdom.

Om det t ex råder halt väglag och en äldre dam faller och bryter benet så kan detta bero på benskörhet, men även på att gångbanan inte var sandad eller på att hon inte hade broddar på skorna. Även om detta inte kommer in direkt i utvärderingsarbetet så kan det ha stor betydelse för vilka slutsatser som ska dras av litteraturgenomgången. Kanske vore det mer ändamålsenligt och kostnadseffektivt att dela ut ”gratis” broddar till alla än att satsa på ökad användning av läkemedel mot benskörhet.

Sociala *konsekvenser* av en ny metod eller ändrad praxis berör t ex patientens möjligheter att leva ett normalt liv när det gäller boende, familjeliv och umgänge, samt vilka resurstillskott och/eller organisationsförändringar som kan krävas för detta. Det kan också handla om patienternas möjligheter att välja sin egen livsstil. Vid utvärdering av medicinska metoder kan sociala aspekter även handla om vilka resurser som krävs för att använda en viss metod och om hur vårdens resurser fördelas i befolkningen [9] (Exempel 12.7).

Exempel 12.7 Sociala konsekvenser av behandlingsmetoder.

Personer med psykisk sjukdom lever ofta ensamma och har svårt att få arbete och försörja sig. Sjukdomen kan dessutom innebära en extra belastning på anhöriga och andra i omgivningen. Behandling av psykisk sjukdom kan därför medföra sociala konsekvenser för de sjuka, liksom för deras anhöriga. Sjukvården har ibland svårt att identifiera andra medicinska problem hos psykiskt sjuka, vilket kan medföra inadekvat behandling av andra hälsoproblem.

Råd om fysisk aktivitet som omfattar förslag om ändrade levnadsvanor riskerar att komma i konflikt med patientens eget ”livsprojekt”.

Ett viktigt syfte med hälso- och sjukvården är att förbättra och/eller återställa förlorad funktions- och arbetsförmåga. I utvärderingen av sociala aspekter bör därför ingå en bedömning av de studerade metodernas effekter i dessa avseenden och även fördelningen av dessa effekter i befolkningen. Det handlar också om konsekvenser av olika åtgärder och beslut om vårdstruktur och vårdutbud, och ligger därmed i vissa avseenden nära de ekonomiska aspekterna. Det finns även släktskap och överlappning med etiska aspekter. Var går gränsen mellan den enskilda individens ansvar för sin hälsa och det ansvar som ligger på anhöriga, arbetsgivare och samhälle? Har alla lika möjligheter att tillgodogöra sig olika slags insatser, såsom förebyggande åtgärder?

Hur ska sociala aspekter beaktas och redovisas?

De sociala aspekterna kan liksom etiska aspekter bli aktuella i olika led av SBU:s projektarbete och i rapporterna. I projektplanen anges hur de sociala aspekterna ska hanteras i projektet. Det kan ofta vara lämpligt att försöka samordna denna del med arbetet rörande etiska aspekter och/eller att följa i princip samma arbetsmodell:

- Inledande analys
- Litteraturgranskning
- Social analys av utvärderingsområdet
- Diskussion med berörda grupper
- Slutsatser
- Förväntade praxisförändringar, implementering.

Inledande analys

I varje projekt ska i samband med formulering av frågeställningarna övervägas *vilka* sociala aspekter som kan och bör belysas inom projektet. Det är alltså angeläget att i ett tidigt skede av arbetet identifiera och beskriva tänkbara sociala frågor och problemområden kring de metoder som ska utvärderas.

Litteraturgranskning

Den litteratur som söks för att utvärdera en metods effekter, biverkningar och kostnadseffektivitet kan också innehålla viktig information som rör det sociala området. Dessa uppgifter används i resonemanget om de sociala aspekterna. Den sociala miljön varierar med kulturella, ekonomiska och sociala villkor. Det innebär att systematisk litteraturgranskning som rör sociala aspekter sannolikt måste avgränsas till studier som omfattar befolkningen i Sverige. I vissa fall kan också studier från övriga nordiska länder bedömas som relevanta. Ett alternativ eller komplement till litteraturgranskning kan vara att göra praxisstudier med fokus på socioekonomiska skillnader.

Social analys av utvärderingsområdet

I den sociala analysen försöker projektgruppen överblicka den händelsekedja som är relaterad till användningen av de utvärderade metoderna. Det kan resultera i ett antal frågor som belyser de sociala aspekterna men som kanske inte alltid behöver besvaras av projektet.

Diskussion med berörda grupper

Liksom för etiska aspekter kan det i vissa fall vara lämpligt att de grupper som kommer att vara berörda av rapporten och de sociala konsekvenserna av resultaten får ta del av och lämna synpunkter på den sociala analysen av utvärderingsområdet innan rapporten publiceras.

Slutsatser

De sociala frågorna och analysen av dessa räknas också som resultat av utvärderingen och bör komma till uttryck i slutsatserna. De kan även vara en del av de viktigaste budskapen i en kommunikationsplan.

Förväntade praxisförändringar, implementering

De sociala konsekvenserna analyseras lättast när utvärderingen av de medicinska effekterna och de ekonomiska konsekvenserna är färdig. Relevant vetenskaplig litteratur, resultatet av praxisundersökningen och projektgruppens samlade kunskaper och kompetens inom ämnesområdet används för att besvara följande frågor:

- Ger vi vård på lika villkor?
- Vilka sociala konsekvenser för arbetslivet eller det sociala livet kan förväntas vid en förändrad praxis?
- Finns resurser i form av personal, pengar etc för att genomföra en eventuell praxisförändring?

Resultatet av diskussionen kan redovisas i Avsnittet ”Konsekvenser av förväntade praxisförändringar” (eller motsvarande).

Referenser

1. Braunack-Mayer AJ. Ethics and health technology assessment: Handmaiden and/or critic? Int J Technol Assess Health Care 2006;22:307-21.
2. Hofmann B. Toward a procedure for integrating moral issues in health technology assessment. Int J Technol Assess Health Care 2005;21:312-8.
3. Beauchamp T, Childress J. The principles of biomedical ethics. Oxford; 2001.
4. Gillon R. Principles of Health Care Ethics. Gillon R, Lloyd A, editors. John Wiley & Sons. Chichester; 1994.
5. Helsingforsdeklarationen. The World Medical Association, http://www.wma.net/en/30publications/10policies/b3/index.html
6. Weingarten MA, Paul M, Leibovici L. Assessing ethics of trials in systematic reviews. BMJ 2004;328:1013-4.
7. Autti-Rämö I, Mäkelä M. Ethical evaluation in health technology assessment reports: An eclectic approach. Int J Technol Assess Health Care 2007;23:1-8.
8. Hermerén G. Riktlinjer för etisk värdering av medicinsk humanforskning. Stockholm: Medicinska forskningsrådet; 2000. MFR-rapport 2. www.infovoice.se/fou/bok/diverse/etik2000.pdf
9. Lehoux P, Williams-Jones B. Mapping the integration of social and ethical issues in health technology assessment. Int J Technol Assess Health Care 2007;23:9-16.

Bilaga 1. **Mall för bedömning av relevans**

VERSION 2010:1.2

Författare: _____ År: _____ Artikelnummer: _____

Relevans	**Ja**	**Nej**	**Oklart**	**Ej tillämpl**
1. Studiepopulation				
a) Är den population som deltagarna togs från tydligt beskriven och relevant?	❑	❑	❑	❑
b) Är sättet att rekrytera deltagare acceptabelt?	❑	❑	❑	❑
c) Är studiens inklusionskriterier adekvata?	❑	❑	❑	❑
d) Är studiens exklusionskriterier adekvata? [1]	❑	❑	❑	❑
2. Undersökt intervention				
a) Är den undersökta interventionen relevant? [2]	❑	❑	❑	❑
b) Är den undersökta interventionen administrerad/utförd på ett korrekt sätt? [3]	❑	❑	❑	❑
c) Är den undersökta interventionen administrerad/utförd på ett reproducerbart sätt? [4]	❑	❑	❑	❑
3. Jämförelseintervention				
a) Är jämförelseinterventionen relevant? [5]	❑	❑	❑	❑
b) Kan man utesluta att val av jämförelse-intervention, dos eller administrationssätt/utförande medfört ett systematiskt fel till förmån för endera interventionen?	❑	❑	❑	❑
4. Effektmått				
a) Har undersökta effektmått klinisk relevans?	❑	❑	❑	❑
5. Studielängd [6]				
a) Är studiens längd adekvat?	❑	❑	❑	❑
b) Är uppföljningstiden adekvat?	❑	❑	❑	❑

Total bedömning av studierelevans	
Relevant ❑	Inte relevant ❑

Kommentarer till mallen för bedömning av relevans

Studiepopulation

1. Valet av exklusionskriterier påverkar ofta studiens generaliserbarhet och kan påverka utfallet. Ofta exkluderas patienter felaktigt på grund av bl a samsjuklighet, ålder, samtidigt intag av vanliga mediciner eller kvinnligt kön. Många andra skäl till varför patienter utesluts har rapporterats. Knappt hälften av de exklusionskriterier som anges i randomiserade studier som publicerats i välrenommerade tidskrifter har rapporterats vara välgrundade.

Undersökt intervention

2. Exempel på interventioner med bristande relevans kan vara t ex när beredningsformen inte är godkänd i Sverige.

3. För läkemedelsstudier finns risker för felaktig dos, administrationssätt, beredningsform, administrationstidpunkt. För metoder som kirurgi och psykoterapi kan liknande resonemang användas (val av teknik, tidpunkt etc).

4. Uppnår alla behandlare samma resultat, eller beror resultatet på behandlarens skicklighet (snarare än själva behandlingen)? Detta kan vara speciellt relevant för psykoterapi, kirurgi och andra manuella tekniker.

Jämförelseintervention

5. Läkemedelsstudier: Har man använt placebo även om det fanns aktiva kontroller att tillgå vid studiens utförande? Är jämförelseinterventionen representativ? Det är t ex vanligt med studier där man använder kontrolläkemedel som visat sig vara sämre än genomsnittet eller inte ens är tillgängliga i Sverige. Se även punkt 2 ovan.

Studielängd

6. Har studien avbrutits i förtid? Varför?

Bilaga 2. **Mall för kvalitetsgranskning av randomiserade studier**

VERSION 2012:2.1

Granskningen av en studie gäller i första hand studiekvalitet, dvs risk för systematiska fel och risk för intressekonflikter (A). I den sammanvägda bedömningen av alla inkluderade studier enligt GRADE inkluderar man också studiernas överensstämmelse (B), överförbarhet (C), precision (D), publikationsbias (E), effektstorlek (F), dos–responssamband (G) och sannolikhet att effekten är underskattad (H).

Författare: _____ År: _____ Artikelnummer: _____

Alternativet ”oklart” används när uppgiften inte går att få fram från texten. Alternativet ”ej tillämpligt” väljs när frågan inte är relevant. Specificera i kommentarsfältet.

A. Granskning av studiens begränsningar – eventuella systematiska fel (bias)	**Ja**	**Nej**	**Oklart**	**Ej tillämpligt**
A1. Selektionsbias				
a) Användes en lämplig randomiseringsmetod?	❑	❑	❑	❑
b) Om studien har använt någon form av begränsning i randomiseringsprocessen (t ex block, strata, minimisering), är skälen till detta adekvata?	❑	❑	❑	❑
c) Var grupperna sammansatta på ett tillräckligt likartat sätt?	❑	❑	❑	❑
d) Om man har korrigerat för obalanser i baslinjevariabler, har det skett på ett adekvat sätt?	❑	❑	❑	❑
Kommentarer:				
Bedömning av risk för selektionsbias:	Låg ❑	Medelhög ❑	Hög ❑	
A2. Behandlingsbias				
a) Var studiedeltagarna blindade?	❑	❑	❑	❑
b) Var behandlare/prövare blindade?	❑	❑	❑	❑
c) Var följsamhet i grupperna acceptabel enligt tillförlitlig dokumentation?	❑	❑	❑	❑
d) Har deltagarna i övrigt behandlats/exponerats på samma sätt bortsett från interventionen?	❑	❑	❑	❑
Kommentarer:				
Bedömning av risk för behandlingsbias:	Låg ❑	Medelhög ❑	Hög ❑	

A. Fortsättning	Ja	Nej	Oklart	Ej tillämpligt
A3. Bedömningsbias (per utfallsmått)				
a) Var utfallsmåttet okänsligt för bedömningsbias?	❑	❑	❑	❑
b) Var de personer som utvärderade resultaten blindade för vilken intervention som gavs?	❑	❑	❑	❑
c) Var personerna som utvärderade utfallet opartiska?	❑	❑	❑	❑
d) Var utfallet definierat på ett lämpligt sätt?	❑	❑	❑	❑
e) Var utfallet identifierat/diagnostiserat med validerade mätmetoder?	❑	❑	❑	❑
f) Har utfallet mätts vid optimala tidpunkter?	❑	❑	❑	❑
g) Var valet av statistiskt mått för rapporterat utfall lämpligt?	❑	❑	❑	❑
h) Var den analyserade populationen (ITT eller PP) lämplig för den fråga som är föremål för studien?	❑	❑	❑	❑
Kommentarer:				
Bedömning av risk för bedömningsbias:	Låg ❑	Medelhög ❑	Hög ❑	
A4. Bortfallsbias (per utfallsmått)				
a) Var bortfallet tillfredsställande lågt i förhållande till populationens storlek?	❑	❑	❑	❑
b) Var bortfallet tillfredsställande lågt i förhållande till storleken på utfallet?	❑	❑	❑	❑
c) Var bortfallets storlek balanserad mellan grupperna?	❑	❑	❑	❑
d) Var relevanta baslinjevariabler balanserade mellan de som avbryter sitt deltagande och de som fullföljer studien?	❑	❑	❑	❑
e) Var den statistiska hanteringen av bortfallet adekvat?	❑	❑	❑	❑
f) Var orsakerna till bortfallet analyserade?	❑	❑	❑	❑
Kommentarer:				
Bedömning av risk för bortfallsbias:	Låg ❑	Medelhög ❑	Hög ❑	

A. Fortsättning	**Ja**	**Nej**	**Oklart**	**Ej tillämpligt**
A5. Rapporteringsbias				
a) Har studien följt ett i förväg publicerat studieprotokoll?	❑	❑	❑	❑
b) Angavs vilket/vilka utfallsmått som var primära respektive sekundära?	❑	❑	❑	❑
c) Redovisades alla i studieprotokollet angivna utfallsmått på ett fullständigt sätt?	❑	❑	❑	❑
d) Mättes biverkningar/komplikationer på ett systematiskt sätt?	❑	❑	❑	❑
e) Redovisades enbart utfallsmått som angivits i förväg i studieprotokollet?	❑	❑	❑	❑
f) Var tidpunkterna för analys angivna i förväg?	❑	❑	❑	❑
Kommentarer:				
Bedömning av risk för rapporteringsbias:	Låg ❑	Medelhög ❑	Hög ❑	
A6. Intressekonflikter				
a) Föreligger, baserat på författarnas angivna bindningar och jäv, låg eller obefintlig risk att studiens resultat har påverkats av intressekonflikter?	❑	❑	❑	❑
b) Föreligger, baserat på uppgifter om studiens finansiering, låg eller obefintlig risk att studien har påverkats av en finansiär med ekonomiskt intresse i resultatet?	❑	❑	❑	❑
c) Föreligger låg eller obefintlig risk för annan form av intressekonflikt (t ex att författarna har utvecklat interventionen)?	❑	❑	❑	❑
Kommentarer:				
Bedömning av risk för intressekonflikt:	Låg ❑	Medelhög ❑	Hög ❑	

Sammanvägning av risk för bias (per utfallsmått)	Låg	Medelhög	Hög
A1. Selektionsbias	❑	❑	❑
A2. Behandlingsbias	❑	❑	❑
A3. Bedömningsbias	❑	❑	❑
A4. Bortfallsbias	❑	❑	❑
A5. Rapporteringsbias	❑	❑	❑
A6. Intressekonfliktbias	❑	❑	❑
Kommentarer:			
Sammanfattande bedömning av risk för systematiska fel (bias):	Låg ❑	Medelhög ❑	Hög ❑

Underlag för sammanvägd bedömning enligt GRADE

B. Bristande överensstämmelse mellan studierna
Hanteras endast på syntesnivå

C. Granskning av studiens överförbarhet	Ja	Nej	Delvis	Ej tillämpligt
a) Överensstämmer sammanhanget och kontrollgruppens villkor med den situation som SBU/HTA-rapportens slutsatser avser?	❑	❑	❑	❑
b) Är den inkluderade studiepopulationen tillräckligt lik den population som SBU/HTA-rapportens slutsatser avser?	❑	❑	❑	❑
c) Är interventionen relevant för de förhållanden som SBU/HTA-rapportens slutsatser avser?	❑	❑	❑	❑
Kommentar:				
Bedömning av brister i överförbarhet:	Inga ❑	Vissa ❑	Stora ❑	

D. Granskning av precision	Ja	Nej	Delvis	Ej tillämpligt
a) Är precisionen acceptabel med hänsyn till antal inkluderade individer och antal händelser (utfall)?	❑	❑	❑	❑
Kommentar:				

E. Granskning av publikationsbias
Hanteras endast på syntesnivå

F. Granskning av effektstorlek	**Ja**	**Nej**	**Delvis**	**Ej tillämpligt**
a) Var effekten stor (t ex RR <0,5 eller >2,0)?	❑	❑	❑	❑
b) Var effekten mycket stor (t ex RR <0,2 eller >5,0)?	❑	❑	❑	❑
Kommentar:				

G. Granskning av dos–responssamband	**Ja**	**Nej**	**Delvis**	**Ej tillämpligt**
a) Finns stöd för ett dos–responssamband mellan exponering och utfall?	❑	❑	❑	❑
Kommentar:				

H. Sannolikhet att effekten är underskattad pga ”confounders”
Inte aktuellt på RCT:er

Mall för kvalitetsgranskning av randomiserade studier: förklaringar

Granskningsmallen avser att ge ett systematiskt underlag med vars hjälp man kan bedöma risken för att ett givet utfall man skattat i en studie systematiskt snedvridits (bias) under forskningsarbetet. Konsekvensen av detta är att utfallet antingen underskattas eller överskattas jämfört med ett ”sant” utfall. Även utfallets riktning kan ha missbedömts.

Syftet med mallen är att skapa ett systematiskt och transparent underlag för att diskutera hur stor risken är att skattade utfall i en enskild studie är systematiskt snedvridna. Någon algoritm för att räkna samman kvalitetspoäng erbjuds alltså inte. När det gäller bedömningsbias (A3) och bortfallsbias (A4) behöver granskningen ske per utfallsmått eftersom kvalitetsbrister/bias kan skilja sig för olika utfallsmått.

För att resultaten ska kunna användas för evidensgradering enligt GRADE krävs ytterligare information i form av sammanställningar på syntesnivå, alltså sammanvägning om fler än en studie finns. I vissa fall kan sammanställningar endast ske på syntesnivå, t ex bristande samstämmighet (”inconsistency”), precision och publikationsbias.

A. Granskning av studiens begränsningar – eventuella systematiska fel

A1. Risk för selektionsbias (”selection bias”)

Med ”selektionsbias” avses systematiska fel som är relaterade till hur studien har hanterat urval av försökspersoner (motsvarande) samt indelning i interventions- och kontrollgrupper.

Risk för selektionsbias föreligger då interventionsgruppen respektive kontrollgruppen inte är tillräckligt lika varandra vid baslinjen avseende kända såväl som okända risk- och skyddsfaktorer. Det utfall man funnit i studien kan då åtminstone delvis bero på dessa skillnader och på så vis snedvrida resultatet. Randomiseringen bör ske på ett oförutsägbart sätt och processen bör inte vara möjlig att manipulera. Detta kan förhindras t ex genom att allokeringen sker med hjälp en datorgenererad slumptalsserie samt genom att processen är maskerad t ex med slutna kuvert.

Ibland begränsas randomisering för att åstadkomma lika stora grupper (t ex blockrandomisering) eller för att skapa balans mellan grupperna avseende sådana egenskaper hos deltagarna som kan påverka resultaten (t ex stratifierad randomisering). Detta kan öka förutsägbarheten avseende vilken grupp en given individ kommer att hamna i. Detta gäller speciellt om blocken är små respektive om varje stratum innehåller få personer.

A1d. Post hoc-justering av utfallet baserat på skillnader i kända baslinjefaktorer är kontroversiellt. Om det används som ett sätt att testa känsligheten av ett positivt utfall är det okej, men för att ändra ett negativt utfall i grundanalysen till ett positivt krävs mycket hållbara argument.

A2. Risk för behandlingsbias (”performance bias”)

Med ”behandlingsbias” avses systematiska fel som är relaterade till hur personer som tillhör interventions- respektive kontrollgruppen har behandlats i studien.

Risk för behandlingsbias föreligger då interventions- eller kontrollgruppen exponeras för något annat än det som jämförelsen syftar till att mäta, t ex annan behandling mot aktuell sjukdom än godkänd standardbehandling. Det utfall man funnit i studien kan då åtminstone delvis bero på dessa skillnader och på så vis snedvrida resultatet.

Om man vill skatta effekten av en given behandling bör kontrollgruppen (placebo- eller obehandlad kontroll) exponeras för exakt samma sak som behandlingsgruppen bortsett från själva behandlingen. Om annat förekommer kan effekten som redovisas i studien överskatta eller underskatta den sanna effekten, detta gäller även effektens riktning, dvs risk för behandlingsbias föreligger.

Om man vill skatta effekten av en behandling jämfört med en alternativ (aktiv) behandling bör ingen av grupperna exponeras för något annat än det som ingår i de båda behandlingarna. Om annat förekommer kan effekten i studien överskatta eller underskatta den sanna effektskillnaden, detta gäller även effektens riktning, dvs risk för behandlingsbias förekommer.

Skillnader kan avse felaktig behandling, ofullständig behandling, behandlingsavbrott, tillägg utanför studieprotokollet m m. Risken för bias kan minska om behandlare och patienter är ovetande om gruppindelningen (blindad studie) och om det finns strukturerad kontroll av implementeringen (t ex en checklista eller en manual).

A2a/b. Det är önskvärt att både patienter, prövare (och utvärderare, se A3b) är blindade i en studie. Ibland kan det av praktiska skäl vara svårt eller omöjligt att dölja för prövare och/eller patient vilken behandling som ges. Blindningen kan också misslyckats pga karakteristiska effekter eller biverkningar av aktiv behandling, exempelvis muntorrhet vid behandling med neuroleptika och underlivsblödningar vid behandling med östrogen. I vissa fall är det möjligt att ge biverkningsmotverkande medel som tillägg till aktiv behandling för att minska risken för att blindningen äventyras. Andra faktorer som kan försvåra blindningen är bristande likhet mellan tabletter, inhalationspreparat etc avseende utseende eller smak. En stor ”placeboeffekt” i kontrollgruppen kan tala för en lyckad blindning. I vissa studier låter man studiedeltagarna gissa om de fått aktiv behandling eller kontroll.

A2c. Kontroll av följsamheten är särskilt viktig då det saknas en signifikant effektskillnad i utfall mellan grupperna. En bristande följsamhet kan minska såväl interventionens effekter som bieffekter. Detta är alltså extra viktigt vid så kallade ”non-inferiority”(”inte sämre än”)-studier men om interventionen har en signifikant effekt är kontroll av följsamheten ofta av mindre betydelse. Undantag är om det var sämre följsamhet i gruppen som fick referensbehandling. Det senare är tänkbart i en placebokontrollerad studie om blindningen varit otillräcklig, alternativt om en referensbehandling har mycket högre frekvens av biverkningar.

A3. Risk för bedömningsbias (”detection bias”)

Med ”bedömningsbias” avses systematiska fel som är relaterade till hur studien har hanterat genomförande av mätningar och analys av resultat.

Risk för bedömningsbias föreligger då det finns skillnader i hur utfallen i interventionsrespektive kontrollgruppen bestäms. Det utfall man funnit i studien kan då åtminstone delvis bero på dessa skillnader och på så vis snedvrida resultatet. Bedömningsbias, och

därmed studiekvaliteten som helhet, kan vara olika för olika utfallsmått i en och samma studie. Bedömning under A3 kan därför behöva göras separat för olika utfallsmått i samma studie.

A3a. Risken för bias ökar ju mer subjektiva inslag som finns i bedömningen av utfallet. Medan överlevnad/död är robusta utfallsmått är symtomskalor och livskvalitetsmätningar mycket känsliga för bias och i princip oanvändbara i oblindade studier.

A3b. Förutom att den som utvärderar studien är blindad är det också viktigt att det framgår av beskrivningen att all resultatbearbetning utfördes innan prövningskoden bröts.

A3c. I randomiserade studier är ju ofta prövare och utvärderare samma personer, men i större högkvalitativa studier finns ibland oberoende kommittéer (DSMB) som tar ställning till och utvärderar utfallet.

A3d. Här handlar det ofta om hur så kallade kompositmått, dvs kombinerade utfallsmått, är sammansatta eller vilken koppling till klinisk relevans som finns för olika surrogatmått.

A3e. Om mätningen sker med hjälp av en standardiserad metod som validerats med avseende på den aktuella populationen minskar risken för bias.

A3f. Val av mättidpunkt för att optimera möjligheten att upptäcka en skillnad i utfall är särskilt viktigt i så kallade ”non-inferiority”-studier.

A3g. De vanligaste mått som används för dikotoma variabler, exempel ja–nej-variabler, är riskkvot (”risk ratio”, RR), oddskvot (”odds ratio”, OR), absolut riskreduktion/riskskillnad (”risk difference”) och ”number needed to treat” (NNT). ”Hazard ratio” (HR) används för att analysera risken över tid. För kontinuerliga variabler används vanligen absolut skillnad i medelvärde (”difference in means”, ”mean difference”) alternativt definieras gränsen för respons och utfallet rapporteras som ”responder rate”. Alla måtten (helst differensen mellan grupperna) ska redovisas med lämpligt precisionssmått, företrädesvis 95 procents konfidensintervall.

A3h. Resultaten kan analyseras enligt ”intention to treat” (ITT) och/eller per protokoll (PP). En ITT-analys innebär att alla personer som randomiserats följs upp inom sin behandlingsarm oavsett om de fått tilldelad behandling eller inte och är oftast den metod som bör användas. Om resultaten är beräknade på annat sätt än med

ITT finns det risk för att behandlingseffekten blir överskattad. ITT-analysen kan kompletteras med en känslighetsanalys enligt ”worst case scenario” där sämsta tänkbara utfall tillskrivs saknade patienter i den grupp som uppvisar bäst effekt och bästa tänkbara utfall tillskrivs saknade patienter i den grupp som uppvisar sämst effekt. Ibland (speciellt ”non-inferiority”-studier) är det viktigt att även en PP-analys redovisas, vilket innebär att bara de som följt hela studieprotokollet ingår i analysen.

A4. Bortfallsbias (”attrition”)

Med ”bortfallsbias” avses systematiska fel som är relaterade till hur studien har hanterat bortfall, dvs personer som har gått med på att delta i en undersökning men som lämnar denna innan den fullbordas.

Risk för bortfallsbias föreligger då det finns skillnader i bortfallet mellan interventions- och kontrollgruppen. Det utfall man funnit i studien kan då åtminstone delvis bero på dessa skillnader och på så vis snedvrida resultatet. Ett generellt stort bortfall, skillnader i bortfallstorlek samt framför allt orsaksskillnader till bortfall ökar risken för bias. Det bortfall som bedöms här avser bortfall efter randomisering. Man kan aldrig räkna med att bortfall är slumpmässigt. Om sammansättningen av personer i bortfallet inte skiljer sig från dem som finns kvar i studien, är dock en bättre situation än om det finns signifikanta skillnader. Nedanstående exempel kan tjäna som *grova* riktvärden:

- litet (<10 %)
- måttligt (10–19 %)
- stort (20–29 %)
- mycket stort (≥30 %). Undersökningen bedöms ofta sakna informationsvärde vilket kan innebära att studien bör exkluderas.

Bortfallet måste också ställas i relation till storleken (och skillnaden) i utfallet. Ju lägre utfall desto större problem även med små bortfall.

Bortfallet kan variera mellan olika tidpunkter i en studie och mellan olika utfallsmått. Bortfallet är ofta större ju längre tid som har gått. Därmed kan behandlingsresultaten från de sista besöken vara av tveksam validitet, medan resultaten från de första besöken kan vara giltiga.

A4e. Vid analys av studier med bortfall används olika så kallade imputeringsmetoder (dvs hur man ersätter missade mätningar, t ex ”last observation carried forward” (LOCF), ”observed cases” (OC) eller interpoleringar). Det är viktigt att utfall med olika imputeringsmetoder redovisas alternativt att man använt den metod

som är minst gynnsam för utfallet (konservativ). Detta kan förvisso göra att storleken på effekten underskattas. I så kallade ”non-inferiority”-studier ska man tvärtom använda den imputeringsmetod som gynnar utfallet eftersom man annars kan komma fram till en felaktig slutsats om frånvaro av effekt/skillnad.

A5. Rapporteringsbias (”reporting bias”)

Med ”rapporteringsbias” avses systematiska fel som är relaterade till hur studien har hanterat rapportering i relation till sitt protokoll.

Det utfall man funnit i studien kan åtminstone delvis bero på att endast vissa resultat rapporteras, medan andra inte rapporteras. Utfallet riskerar då att såväl överskattas som underskattas. Även utfallets riktning kan ha påverkats.

A5a. Det är inte ovanligt att studier med *negativa* resultat inkluderar förklarande efteranalyser (”explanatory”- eller post hoc-analyser) för att t ex finna vissa subgrupper inom den studerade patientgruppen som kan ha nytta av behandlingen. Dessa analyser kan fylla en viktig hypotesgenererande funktion, men slutsatserna i en *negativ* studie får aldrig baseras på sådana analyser. När en studie visar ett *positivt* utfall för sitt primära utfallsmått är däremot subgruppsanalyser av stort värde för att bedöma generaliserbarheten av resultatet.

A5c/d. Även om redovisade utfallsmått är rimliga, definierade i förväg och adekvat rapporterade kan det finnas andra viktiga utfallsmått som utelämnats. Oftast gäller det utfallsmått för att bedöma biverkningar/risker.

A5f. Det är viktigt att inte fler analyser av studien än vad som angetts i protokollet (och den statistiska planen medger) gjorts. Det är också viktigt att det framgår om den redovisade analysen är en slutanalys eller en förplanerad interimsanalys. Ad hoc interimsanalyser är självklart mycket problematiska speciellt i öppna studier där de kan misstänkas vara datadrivna.

A6. Intressekonfliktbias (”other considerations”)

Om författare till studien kan vinna något på ett givet resultat, så kan detta medföra en överskattning eller underskattning av effekten i den riktning som författaren skulle vinna på.

Sammanvägning

För att dimensionen studiebegränsningar ska kunna beaktas när ett betyg sätts för ett sammanvägt utfallsmått med hjälp av GRADE, krävs att alla ovanstående former av risk för bias vägs samman. Detta sker med fördel i diskussion i expertgrupp.

B. Bristande överensstämmelse mellan studierna (”heterogeneity”)

Hanteras på syntesnivå.

C. Bristande överförbarhet (”indirectness of evidence”)

Med ”överförbarhet” avses möjligheten att tillämpa studiens upplägg, diskussion och resultat på de förhållanden som SBU/HTA-rapporten avser.

Om population, intervention, kontrollalternativ eller utfallsmått avviker från dem som specificerats i SBU/HTA-översikten föreligger överförbarhetsproblem. Det utfall man funnit i studien kan då åtminstone delvis avvika från det ”sanna” utfallet med avseende på hur population, intervention, kontrollalternativ eller utfallsmått har specificerats i översikten. Utfallet kan alltså underskattas såväl som överskattas, vilket även gäller utfallets riktning.

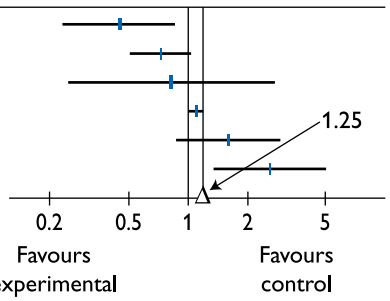
Det är betydligt viktigare att studiepopulationen motsvarar den population man vill dra slutsatser om i SBU/HTA-rapporten, än om studiepopulationen inte är representativ med avseende på syftet i den enskilda studien (t ex beroende på bortfall före randomiseringen).

För att dimensionen överförbarhet ska kunna beaktas när ett betyg sätts med hjälp av GRADE för ett sammanvägt utfallsmått, krävs att ingående studier beaktas som en helhet.

D. Bristande precision (”imprecision”)

Här beaktas två aspekter av precision. För det första, om syftet är att testa om interventionen är bättre än kontrollvillkoret räcker det här med att studera om konfidensintervallet täcker linjen för ”ingen skillnad” (”1” vid binära utfallsmått samt ”0” vid kontinuerliga utfallsmått). Täcks denna linje är precisionen bristande. Resultaten i ”Superior”, ”Non-inferior B” och ”Inferior” har god precision i detta avseende (Figur B2.1). För det andra, om syftet är att testa huruvida interventionen inte är sämre än kontrollinterventionen (ofta rörande biverkningar), krävs även en i förväg kliniskt definierad gräns för hur mycket sämre interventionen får vara utan att det är ett problem (”suggested appreciable harm”, kliniskt relevant skillnad). Om konfidensintervallet inte täcker denna gräns är precisionen god och man kan då dra slutsatsen att interventionen inte var sämre än kontrollinterventionen. I Figur B2.1 har gränsen satts till 1,25. Tre exempel på resultat som kan illustrera detta är ”Superior”, ”Non-inferior A” samt ”Non-inferior B”. Exempel på dålig precision utgörs av ”Imprecise A” och ”Imprecise B”. Observera att datakvaliteten är viktig vid bedömning av precisionen i ”non-inferiority”-utfall. Exempelvis kan en dålig rapportering av biverkningar göra att resultatet ser ut att vara lika i båda behandlingsarmarna.

Study or subgroup	Experimental Events	Total	Control Events	Total	Odds Ratio IV, Fixed, 95% CI
Superior	15	179	30	179	0.45 [0.24, 0.88]
Non-inferior A	200	300	220	300	0.73 [0.51, 1.03]
Imprecise A	5	179	6	179	0.83 [0.25, 2.77]
Non-inferior B	2 000	4 000	1 900	4 000	1.11 [1.01, 1.21]
Imprecise B	30	179	20	179	1.60 [0.87, 2.94]
Inferior	30	156	15	179	2.60 [1.34, 5.05]



Figur B2.1 Illustration av olika tester med skogsdiagram (”forest plot”).

Finns det fler studier som är lämpliga att väga samman ska det sammanvägda konfidensintervallet beaktas.

E. Publikationsbias

Hanteras endast på syntesnivå.

F. Effektstorlek

Hanteras i första hand på syntesnivå. Om ingående studiers kvalitet har föranlett nedgradering kan uppgradering för effektstorlek komma ifråga endast efter noggrann övervägning.

G. Dos–responssamband

Sammanvägs på syntesnivå. Av praktiska skäl kan det vara bra att notera resultatet för den enskilda studien i granskningsmallen.

H. Sannolikhet att effekten är underskattad pga ”confounders”

Inte aktuellt på RCT:er.

Bilaga 3. **Mall för kvalitetsgranskning av observationsstudier**

VERSION 2012:2.1

Granskningen av en studie gäller i första hand studiekvalitet, dvs risk för systematiska fel och risk för intressekonflikter (A). I den sammanvägda bedömningen av alla inkluderade studier enligt GRADE inkluderar man också studiernas överensstämmelse (B), överförbarhet (C), precision (D), publikationsbias (E), effektstorlek (F), dos–responssamband (G) och sannolikhet att effekten är underskattad (H).

Författare: _____ År: _____ Artikelnummer: _____

Alternativet ”oklart” används när uppgiften inte går att få fram från texten. Alternativet ”ej tillämpligt” väljs när frågan inte är relevant.

A. Granskning av studiens begränsningar – eventuella systematiska fel (bias)	**Ja**	**Nej**	**Oklart**	**Ej tillämpligt**
A1. Selektionsbias				
a) Var de observerade grupperna rekryterade på ett tillräckligt likartat sätt?	❑	❑	❑	❑
b) Var de jämförda gruppernas sammansättning tillräckligt lika vid studiestart?	❑	❑	❑	❑
c) Har korrigering av obalanser i baslinjevariabler mellan grupper med olika exponering/behandling gjorts på ett adekvat sätt i den statistiska analysen?	❑	❑	❑	❑
Kommentarer:				
Bedömning av risk för selektionsbias:	Låg ❑	Medelhög ❑	Hög ❑	
A2. Behandlingsbias				
a) Var villkoren (utöver den behandling eller exponering som studerades) för grupperna under behandlings-/exponeringstiden tillräckligt likartade?	❑	❑	❑	❑
b) Var följsamhet gentemot behandling/exponering acceptabel i grupperna?	❑	❑	❑	❑
Kommentarer:				
Bedömning av risk för behandlingsbias:	Låg ❑	Medelhög ❑	Hög ❑	

A. Fortsättning	Ja	Nej	Oklart	Ej tillämpligt
A3. Bedömningsbias (per utfallsmått)				
a) Var utfallsmåttet okänsligt för bedömningsbias?	❑	❑	❑	❑
b) Var personerna som utvärderade utfallet *blindade* för studiedeltagarnas exponeringsstatus?	❑	❑	❑	❑
c) Var personerna som utvärderade utfallet *opartiska?*	❑	❑	❑	❑
d) Var utfallet definierat på ett lämpligt sätt?	❑	❑	❑	❑
e) Mättes utfallet på ett adekvat sätt med standardiserade/definierade mätmetoder?	❑	❑	❑	❑
f) Mättes utfallet på ett adekvat sätt med validerade mätmetoder?	❑	❑	❑	❑
g) Har variationer i exponering över tid tagits med i analysen?	❑	❑	❑	❑
h) Har utfallet mätts vid optimal(a) tidpunkt(er)?	❑	❑	❑	❑
i) Var observatörsöverensstämmelsen acceptabel?	❑	❑	❑	❑
j) Har studien tillämpat ett lämpligt statistiskt mått för rapporterad effekt/samband?	❑	❑	❑	❑
Kommentarer:				
Bedömning av risk för bedömningsbias:	Låg ❑	Medelhög ❑	Hög ❑	
A4. Bortfallsbias (per utfallsmått)				
a) Var bortfallet tillfredsställande lågt i förhållande till populationens storlek?	❑	❑	❑	❑
b) Var bortfallet lika stort inom grupperna?	❑	❑	❑	❑
c) Var relevanta baslinjevariabler lika fördelade mellan bortfallen i interventions- och kontrollgruppen alternativt mellan olika exponeringsgrupper?	❑	❑	❑	❑
d) Var relevanta baslinjevariabler lika fördelade mellan analys- och bortfallgruppen?	❑	❑	❑	❑
e) Var den statistiska hanteringen av bortfallet adekvat?	❑	❑	❑	❑
Kommentarer:				
Bedömning av risk för bortfallsbias:	Låg ❑	Medelhög ❑	Hög ❑	

A. Fortsättning	**Ja**	**Nej**	**Oklart**	**Ej tillämpligt**
A5. Rapporteringsbias				
a) Följde studien ett i förväg fastlagt studieprotokoll?	❑	❑	❑	❑
b) Var utfallsmåtten relevanta?	❑	❑	❑	❑
c) Mättes biverkningar/komplikationer på ett systematiskt sätt?	❑	❑	❑	❑
d) Var tidpunkterna för rapporterad analys relevanta?	❑	❑	❑	❑
Kommentarer:				
Bedömning av risk för rapporteringsbias:	Låg ❑	Medelhög ❑	Hög ❑	
A6. Intressekonflikter				
a) Föreligger, baserat på författarnas angivna bindningar och jäv, låg eller obefintlig risk att studiens resultat har påverkats av intressekonflikter?	❑	❑	❑	❑
b) Föreligger, baserat på uppgifter om studiens finansiering, låg eller obefintlig risk att studien har påverkats av en finansiär med ekonomiskt intresse i resultatet?	❑	❑	❑	❑
c) Föreligger låg eller obefintlig risk för annan form av intressekonflikt (t ex att författarna har utvecklat interventionen)?	❑	❑	❑	❑
Kommentarer:				
Bedömning av risk för intressekonflikt:	Låg ❑	Medelhög ❑	Hög ❑	

Sammanvägning av risk för bias (per utfallsmått)	**Låg**	**Medelhög**	**Hög**
A1. Selektionsbias	❑	❑	❑
A2. Behandlingsbias	❑	❑	❑
A3. Bedömningsbias	❑	❑	❑
A4. Bortfallsbias	❑	❑	❑
A5. Rapporteringsbias	❑	❑	❑
A6. Intressekonfliktbias	❑	❑	❑
Kommentarer:			
Sammanfattande bedömning av risk för systematiska fel (bias):	Låg ❑	Medelhög ❑	Hög ❑

Underlag för sammanvägd bedömning enligt GRADE

B. Bristande överensstämmelse mellan studierna
Hanteras endast på syntesnivå

C. Granskning av studiens överförbarhet	**Ja**	**Nej**	**Delvis**	**Ej tillämpligt**
a) Överensstämmer sammanhanget och kontrollgruppens villkor med den situation som SBU/HTA-rapportens slutsatser avser?	❑	❑	❑	❑
b) Är den inkluderade studiepopulationen tillräckligt lik den population som SBU/HTA-rapportens slutsatser avser?	❑	❑	❑	❑
c) Är interventionen relevant för de förhållanden som SBU/HTA-rapportens slutsatser avser?	❑	❑	❑	❑
Kommentar:				
Bedömning av brister i överförbarhet:	Inga ❑		Vissa ❑	Stora ❑

D. Granskning av precision	**Ja**	**Nej**	**Delvis**	**Ej tillämpligt**
a) Är precisionen acceptabel med hänsyn till antal inkluderade individer och antal händelser (utfall)?	❑	❑	❑	❑
Kommentar:				

E. Granskning av publikationsbias
Hanteras endast på syntesnivå

F. Granskning av effektstorlek	**Ja**	**Nej**	**Delvis**	**Ej tillämpligt**
a) Var effekten stor (t ex RR <0,5 eller >2,0)?	❑	❑	❑	❑
b) Var effekten mycket stor (t ex RR <0,2 eller >5,0)?	❑	❑	❑	❑
Kommentar:				

G. Granskning av dos–responssamband	**Ja**	**Nej**	**Delvis**	**Ej tillämpligt**
a) Finns stöd för ett dos–responssamband mellan exponering och utfall?	❑	❑	❑	❑
Kommentar:				

H. Sannolikhet att effekten är underskattad pga ”confounders”	**Ja**	**Nej**	**Delvis**	**Ej tillämpligt**
Vid enstaka tillfällen kan evidensstyrkan höjas om det är mycket sannolikt att effekten är underskattad.				
a) Finns det starkt stöd för att ”confounders” som studien inte kunnat ta hänsyn till skulle stärka sambandet?	❑	❑	❑	❑
Kommentar:				

Mall för kvalitetsgranskning av observationsstudier: förklaringar

Mallen är i första hand tänkt att användas för granskning av studiekvalitet i prospektiva kohortstudier (del A). I den mån retrospektiva kohortstudier med historiska kontroller, retrospektiva fallserier, tvärsnittsstudier eller andra icke-randomiserade studietyper är aktuella att använda kan mallen användas för dessa med vissa tillägg/anpassningar. Granskningsmallen avser att ge ett systematiskt underlag till stöd för att bedöma risken för att en given effekt i en studie systematiskt snedvridits (bias) under forskningsarbetet. Konsekvensen av detta är att effekten antingen underskattas eller överskattas jämfört med en ”sann” effekt. Även effektens riktning kan ha missbedömts.

Syftet med mallen är att skapa ett systematiskt och transparent underlag för att diskutera hur stor risken är att skattade effektmått/samband i en enskild studie är systematiskt snedvridna. Någon algoritm för att räkna samman kvalitetspoäng erbjuds alltså inte.

För att resultaten ska kunna användas för evidensgradering enligt GRADE krävs ytterligare information i form av sammanställningar av samtliga ingående studier. Det gäller bristande överensstämmelse mellan studierna (B), studiens överförbarhet (C), precision (D), risk för publikationsbias (E), effektstorlek (F), dos–responssamband (G) och granskning av sannolikhet att effekten är underskattad (H). Denna sammanvägning sker vid ett senare tillfälle, men det kan vara lämpligt att vid läsningen av en enskild studie samtidigt kommentera dessa faktorer.

A. Granskning av studiens begränsningar – eventuella systematiska fel

A1. Risk för selektionsbias (”selection bias”)

Med ”selektionsbias” avses systematiska fel som är relaterade till hur studien har hanterat urval av försökspersoner (motsvarande) samt indelning i interventions- och kontrollgrupper.

Risk för selektionsbias kan föreligga då interventionsgruppen inte är tillräckligt lik kontrollgruppen vid baslinjen. Kända, såväl som okända, risk- och skyddsfaktorer bör vara tillräckligt lika i de båda grupperna för att inte snedvrida resultatet. Några viktiga förväxlingsfaktorer (”confounders”) är ålder, kön, bakomliggande sjukdomshistoria och samsjuklighet. En annan viktig faktor är socioekonomi, som sannolikt är den starkaste riskfaktorn för sjuklighet och för tidig död. Genom att använda statistiska metoder som matchning, stratifiering, multivariat regressionsanalys eller ”propensity score”-metodik kan man dock korrigera för kända förväxlingsfaktorer (se A1c).

Risken för selektionsbias är också hög om den åtgärd som studeras är särskilt lämplig att sätta in på vissa försökspersoner som har en särskilt hög eller låg chans för att svara väl på åtgärden.

A1a. Finns det en klar definition av jämförelsegruppen? Har de jämförda grupperna rekryterats på så pass likvärdiga sätt att resultaten inte har snedvridits? Har jämförelsegruppen hämtats från den allmänna befolkningen eller från ett begränsat urval? Om jämförelsegruppen är en historisk kontrollgrupp finns det anledning att vara särskilt försiktig vid värderingen. En viktig fråga är om samma metodik användes för att rekrytera till interventions- respektive jämförelsegrupp.

A1b. Uppgifter som kan visa väsentliga skillnader mellan grupperna finns ofta i en inledande tabell eller under bakgrundsdata (”baseline characteristics”).

A1c. Metoder som kan användas i detta sammanhang är matchning/restriktion, stratifierad analys, multivariat modellanalys (t ex regressionsanalys) eller ”propensity score”-metodik.

Eftersom observationsstudier (2+) i GRADE-systemet genom sin studiedesign redan från början antas ha större risk för selektionsbias än randomiserade studier så måste riskerna för selektionsbias vara mycket stora för att bedömas som höga i detta avsnitt.

A2. Risk för behandlingsbias (”performance bias”)

Med ”behandlingsbias” avses systematiska fel som är relaterade till hur studien har behandlat personer som tillhör interventionsgruppen, respektive jämförelsegruppen.

Risk för behandlingsbias föreligger då interventions- eller kontrollgruppen exponeras för något annat än det som jämförelsen syftar till att mäta. Den effekt studien funnit kan då åtminstone delvis bero på dessa skillnader och på så vis snedvrida resultatet. Skillna-

der kan t ex avse felaktig behandling, ofullständig behandling, behandlingsavbrott eller tillägg utanför studieprotokollet. Risken för systematiska fel kan minska om det finns strukturerad kontroll av implementeringen (t ex en checklista eller en manual).

A2a. Om studien syftar till att skatta effekten av en given behandling/riskfaktor (eventuellt i relation till alternativ behandling) bör kontrollgruppen exponeras för exakt samma sak som behandlingsgruppen bortsett från själva behandlingen. Om annat förekommer kan effekten överskattas eller underskattas. Detta gäller även effektens riktning, dvs risk för behandlingsbias föreligger. Finns det t ex socioekonomiska skillnader mellan behandlings- och kontrollgrupp? Risken är särskilt stor när det gäller preventiva eller symtomlindrande åtgärder som olika grupper av välinformerade individer kan efterfråga vilket kan göra att effekten/risken underskattas. Även om grupperna exponeras för olika faktorer som kan påverka utfallet på ett likvärdigt sätt, kan detta minska studiens känslighet för att upptäcka alternativt utesluta en effekt eller ett risksamband.

A2b. Kontroll av följsamhet gentemot behandling alternativt av exponeringen är fundamental för trovärdigheten i uppnådda resultaten. Speciellt viktigt är detta i de fall resultatet pekar mot avsaknad av effekt/samband vilket kan bero på avsaknad av exponering för riskfaktorn eller interventionen. Den yttersta formen av låg följsamhet är avbrott av behandling eller exponering. Avbrott innebär att försökspersonen avbryter behandlingen eller avslutar exponeringen (utan att ha uppnått det studerade utfallet), men inte nödvändigtvis avbryter uppföljningen (= bortfall, se A4). Det är viktigt att kontrollera:

a) totala andelen avbrott
b) skillnaden i andelen avbrott mellan grupperna
c) skillnader i orsak till avbrott mellan grupperna.

A3. Risk för bedömningsbias (”detection bias”)

Med ”bedömningsbias” avses systematiska fel som är relaterade till hur man i studien har hanterat mätningar av utfall och analys av resultat.

Risk för bedömningsbias föreligger då det finns skillnader i hur utfallen i interventions- respektive kontrollgruppen bestäms. Den effekt studien funnit kan då åtminstone delvis bero på dessa skillnader och snedvrida resultatet. Bedömningsbias, och därmed studiekvaliteten som helhet, kan vara olika för olika utfallsmått i en och samma studie. Bedömning under A3 kan därför behöva ske separat för olika utfallsmått i samma studie.

Svaret på vissa av delfrågorna kan göra att andra delfrågor blir mindre eller inte alls relevanta. Exempelvis är relevansen för frågor om blindning (A3b) beroende av hur robust utfallsvariabeln är (A3a).

A3a. Risken för systematiska fel ökar ju mer subjektiva inslag som finns i bedömningen av utfallet. Medan överlevnad/död är robusta utfallsmått är symtomskalor och livskvalitetsmätningar mycket känsliga för systematiska fel.

A3b. Om personerna som mäter utfallen (patologen, röntgenologen, psykologen) eller som utvärderar resultatet av mätningen (”forskaren”) känner till vilka försökspersoner som fått en viss behandling/exponering kan det öka risken för systematiska fel.

A3c. Om samma personal som deltar i behandling eller i studiens genomförande också bedömer utfallet ökar risken för systematiska fel. Vid tillfredsställande blindning är dock opartiskhet av liten betydelse.

A3d. Här handlar det ofta om hur så kallade kompositmått, dvs kombinerade effektmått, är sammansatta eller vilken koppling olika surrogatmått har till klinisk relevans. Vid negativt utfall är det viktigt att det valda utfallsmåttet är tillräckligt känsligt och att konfidensintervallet är tillräckligt smalt för att det ska vara möjligt att utesluta en effekt av klinisk relevant storlek.

A3e/f. Risken för systematiska fel minskar om mätningen sker med hjälp av en standardiserad eller definierad metod som validerats med avseende på den aktuella populationen.

A3g. Möjligheten att upptäcka (liksom att utesluta) effekter/samband ökar om exponeringen uppskattats vid upprepade (optimala) tidpunkter under studien.

A3h. Felaktigt val av tidpunkt för mätning kan göra att utfallet underskattas. Detta är särskilt viktigt vid ”non-inferiority”(”inte sämre än”)-studier eller då slutsatsen är att effekt saknas.

A3i. Vid utfallsregistrering kan observatörsvariationen vara en svaghet. Ett exempel är om flera observatörer ska utvärdera röntgenbilder eller cytologiska prov. Då ska observatörsöverensstämmelse mellan alla, eller ett större antal av observatörerna, vara rapporterat. Detta kan ske i form av kappa-överensstämmelse eller ”intraclass correlation coefficient” (ICC), beroende på vilken skala som använts.

A3j. De vanligaste mått som används för dikotoma variabler (t ex ja–nej-variabler) är:

- riskkvot (”risk ratio”, RR),
- oddskvot (”odds ratio”, OR),
- absolut riskreduktion/riskskillnad (”risk difference”), och
- ”number needed to treat” (NNT).

”Hazard ratio” (HR) används för att analysera risken över tid.

För kontinuerliga variabler används vanligen absolut skillnad i medelvärde (”difference in means”, ”mean difference”), standardiserad medelskillnad (”standardized mean difference” ≈ ”Cohen’s d” ≈ ”Hedges g”), alternativt definieras gränsen för respons och utfallet rapporteras som ”responder rate”. Vid sådan dikotomisering av kontinuerliga variabler är det viktigt att intervallens gräns(er) motiverats trovärdigt eller är ”gängse”.

Alla måtten (helst differensen mellan grupperna) ska redovisas med lämpligt precisionsmått, företrädesvis 95 procents konfidensintervall. Bedöm om konfidensintervall eller andra relevanta mått redovisas på ett adekvat sätt eller om det finns en motivering för att sådana uppgifter saknas. Det kan t ex gälla vid totalundersökningar av stora datamaterial.

A4. Bortfallsbias (”attrition”)

Med ”bortfallsbias” avses systematiska fel som är relaterade till hur studien har hanterat bortfall, dvs personer som har gått med på att delta i en undersökning men som lämnar denna innan deras medverkan/uppföljning är klar. Den engelska termen är ”loss to follow-up”.

Risk för bortfallsbias föreligger då det finns skillnader i bortfallet mellan interventions- och kontrollgruppen. Den effekt studien funnit kan då åtminstone delvis bero på dessa skillnader och på så vis snedvrida resultatet. Ett generellt stort bortfall, skillnader i bortfallstorlek samt framför allt orsaksskillnader till bortfall ökar risken för systematiska fel. Det bortfall som bedöms här avser bortfall efter inklusion i studien. Man kan aldrig räkna med att bortfall är slumpmässigt.

Stora bortfall ökar generellt sett risken för att resultaten kan vara påverkade av systematiska fel. Bortfallet kan variera mellan olika tidpunkter och olika effektmått. Bortfallsanalysen görs därför separat för de aktuella utfallen. Vid långtidsuppföljning kan man få acceptera något högre bortfall.

Bortfallet kan variera mellan olika tidpunkter i en studie och mellan olika effektmått. Bortfallet är ofta större ju längre tid som har gått. Därmed kan behandlingsresultaten från de sista mättillfällena vara av tveksam validitet, medan resultaten från de första mättillfällena kan vara giltiga.

A4a. Som ett riktvärde för läkemedelsstudier är risken liten om bortfallet är mindre än 10 procent, medelstor om bortfallet ligger mellan 10 och 19 procent och stor om bortfallet är mellan 20 och 29 procent. Om bortfallet i läkemedelsstudier är 30 procent eller mer är informationsvärdet tveksamt och studien kan eventuellt sorteras bort. Notera att andra värden kan gälla om det inte är läkemedelsstudier. Bortfallet måste också ställas i relation till storleken (och skillnaden) i utfallet. Ju lägre utfall desto större problem även med små bortfall.

A4c. Skillnader i baslinjevariabler hos bortfall i interventions- och kontrollgrupp alternativt grupper med olika exponering för riskfaktorer är allvarliga eftersom de kan snedvrida utfallet särskilt om det rör baslinjefaktorer med direkt koppling till utfallet (t ex sjukdomsstadium vid överlevnadsutfall).

A4d. Om sammansättningen av personer i bortfallet skiljer från dem som finns kvar i studien, kan det påverka studiens möjlighet att upptäcka relevanta effekter och överförbarhet (t ex att patienter med progredierande sjukdom inte orkar fylla i livskvalitetsfrågeformulär).

A4e. Vid analys av studier med bortfall används olika så kallade imputeringsmetoder (dvs hur man ersätter missade mätningar, t ex ”last observation carried forward” (LOCF), ”observed cases” (OC) eller interpoleringar). Det är viktigt att utfall med olika imputeringsmetoder redovisas alternativt att man använt den metod som är minst gynnsam för utfallet (konservativ). Detta kan förvisso göra att storleken på effekten underskattas. I så kallade ”non-inferiority”-studier ska man tvärtom använda den imputeringsmetod som gynnar utfallet eftersom man annars kan komma fram till en felaktig slutsats om frånvaro av effekt/skillnad.

A5. Rapporteringsbias (”reporting bias”)

Med ”rapporteringssbias” avses systematiska fel som är relaterade till hur studien har hanterat protokoll och rapportering.

Den effekt studien funnit kan åtminstone delvis bero på att vissa resultat rapporteras, medan andra inte rapporteras. Effekten riskerar då att såväl överskattas som underskattas. Även effektens riktning kan ha påverkats.

A5a. Tillgång till studiens protokoll är av stort värde för att bedöma betydelsen av rapporterade fynd eftersom det inte är ovanligt att studier med *negativa* resultat inkluderar förklarande efteranalyser (”explanatory”- eller post hoc-analyser) för att t ex finna vissa subgrupper inom den studerade patientgruppen som kan ha nytta av behandlingen alternativt hos vilka samband identifieras. Dessa analyser kan fylla en viktig hypotesgenererande funktion, men slutsatserna i en primärt *negativ* studie får aldrig baseras på sådana subgruppsanalyser. När en studie visar ett *positivt* utfall för sitt primära utfallsmått är däremot subgruppsanalyser av stort värde för att bedöma generaliserbarheten av resultatet.

A5b. Det är viktigt att det går att klargöra vilka utfall som mätts, analyserats respektive rapporterats. Utfall som mätts eller analyserats men inte rapporterats och därför inte tagits hänsyn till i den statistiska analysen gör att betydelsen av interventionen/sambandet kan missbedömas.

A5c. En undermålig rapportering av risker med en intervention riskerar att överskatta dess ändamålsenlighet (nytta/risk).

A5d. Olika effekter kan ha mätts vid upprepade tillfällen men det är viktigt att det inte gjorts fler *analyser* av studien än vad som angetts i protokollet (och den statistiska planen medger). Det är också viktigt att det framgår om den redovisade analysen är en slutanalys eller en förplanerad interimsanalys. Ad hoc interimsanalyser är självklart mycket problematiska speciellt i öppna studier där de kan misstänkas vara datadrivna. Även ”lege artis” interimsanalyser riskerar att överskatta effekter av en intervention. När det gäller studier som inte påvisar någon effekt av en intervention är det viktigt att tidpunkten för analys är optimalt vald för att kunna påvisa en möjlig effekt.

A6. Intressekonflikter (”other considerations”)

Om författare till studien kan vinna något på ett givet resultat, kan detta medföra en överskattning eller underskattning av effekten i den riktning som författaren skulle vinna på. Det kan t ex vara problematiskt om författarna själva har utvecklat den intervention som studerades.

Sammanvägning

För att bedöma den sammanvägda evidensen med hjälp av GRADE, krävs att även nedanstående faktorer vägs samman i en slutgiltig bedömning.

B. Bristande överensstämmelse mellan studierna (”heterogeneity”)

Görs t ex om möjligt i form av metaanalyser eller liknande, men utgår i granskning av enskild studie.

C. Bristande överförbarhet (”indirectness of evidence”)

Med ”överförbarhet” avses möjligheten att tillämpa studiens upplägg, diskussion och resultat i svenska förhållanden.

Om population, intervention, kontrollalternativ eller effektmått i en studie avviker från dem som specificerats som adekvat för svenska förhållanden och därmed frågeställningen i översikten, så föreligger överförbarhetsproblem. Den effekt studien funnit kan då åtminstone delvis avvika från den ”sanna” effekten med avseende på hur population, intervention, kontrollgruppens villkor eller effektmått som specificerats i översikten. Effekten kan alltså såväl underskattas som överskattas vad gäller svenska förhållanden.

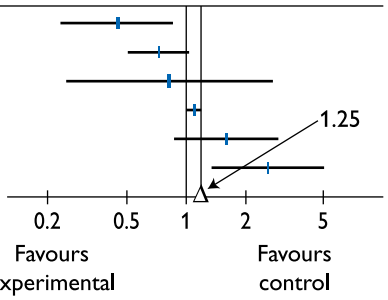
Det är viktigt att populationen i studien motsvarar den population som är aktuell i SBU/HTA-rapporten.

För att kunna bedöma överförbarhet med hjälp av GRADE för ett sammanvägt effektmått krävs att de ingående studierna beaktas som en helhet.

D. Bristande precision (”imprecision”)

Här beaktas två aspekter av precision. För det första, om syftet är att testa om interventionen är bättre än kontrollvillkoret så räcker det här med att studera om konfidensintervallet täcker linjen för ”ingen skillnad” (”1” vid binära utfallsmått samt ”0” vid kontinuerliga utfallsmått). Täcks denna linje är precisionen bristande. Resultaten i ”Superior”, ”Non-inferior B” och ”Inferior” har god precision i detta avseende (Figur B3.1). För det andra, om syftet är att testa huruvida interventionen inte är sämre än kontrollinterventionen (ofta rörande biverkningar), krävs även en i förväg

Study or subgroup	Experimental Events	Experimental Total	Control Events	Control Total	Odds Ratio IV, Fixed, 95% CI
Superior	15	179	30	179	0.45 [0.24, 0.88]
Non-inferior A	200	300	220	300	0.73 [0.51, 1.03]
Imprecise A	5	179	6	179	0.83 [0.25, 2.77]
Non-inferior B	2 000	4 000	1 900	4 000	1.11 [1.01, 1.21]
Imprecise B	30	179	20	179	1.60 [0.87, 2.94]
Inferior	30	156	15	179	2.60 [1.34, 5.05]



Figur B3.1 Illustration av olika tester med skogsdiagram (”forest plot”).

kliniskt definierad gräns för hur mycket sämre interventionen får vara utan att det är ett problem (”suggested appreciable harm”, kliniskt relevant skillnad). Om konfidensintervallet inte täcker denna gräns är precisionen god och man kan då dra slutsatsen att interventionen inte var sämre än kontrollgruppen. I Figur B3.1 har gränsen satts till 1,25. Tre exempel på resultat som kan illustrera detta är ”Superior”, ”Non-inferior A” samt ”Non-inferior B”. Exempel på dålig precision utgörs av ”Imprecise A” och ”Imprecise B”. Observera att datakvaliteten är viktig vid bedömning av precisionen i ”non-inferiority”-utfall. Exempelvis kan en dålig rapportering av biverkningar göra att resultatet ser ut att vara lika i båda behandlingsarmarna.

Finns det fler studier som är lämpliga att väga samman ska det sammanvägda konfidensintervallet beaktas.

E. Publikationsbias

Sammanvägs på syntesnivå.

F. Effektstorlek

Sammanvägs på syntesnivå. Av praktiska skäl kan det vara bra att notera resultatet för den enskilda studien i granskningsmallen.

G. Dos–responssamband

Slutgiltiga bedömningen hanteras på syntesnivå.

H. Sannolikhet att effekten är underskattad pga ”confounders”

Sammanvägs på syntesnivå. Av praktiska skäl kan det vara bra att notera resultatet för den enskilda studien i granskningsmallen. Vid enstaka tillfällen kan evidensstyrkan justeras upp om det är mycket sannolikt att studierna underskattat effekten. Det kan gälla när ”confounders” som studien inte kunnat justera för, talar för att effekten är underskattad.

Bilaga 4. **Mall för kvalitetsgranskning av diagnostiska studier (QUADAS)** [1,2]

Författare: _____ År: _____ Artikelnummer: _____

Mallen består av 11 enskilda kriterier [2]. Hur olika typer av bias kan påverka resultat visas i Tabell 7.2 i SBU:s handbok och i förklaring/kommentarer.

	Ja	Nej	Oklart
1. Var sammansättningen av patientgruppen (spektrum) representativ för de patienter som kommer att få testet i praktiken? *Undvikande av spektrumbias*	❑	❑	❑
2. Är det troligt att referenstestet korrekt klassificerar det sökta tillståndet? *Undvikande av felklassifikationsbias*	❑	❑	❑
3. Var tidsintervallet mellan referenstest och indextest så kort att det studerade tillståndet inte kunnat förändras mellan de båda testen? (Acceptabel fördröjning mellan testerna) *Undvikande av sjukdomsprogressionsbias*	❑	❑	❑
4. Genomgick samtliga patienter eller ett slumpmässigt urval av patienter det avsedda referenstestet? *Undvikande av partiell verifikationsbias*	❑	❑	❑
5. Användes ett och samma referenstest oberoende av vilket resultat som erhölls på indextestet? *Undvikande av differentiell verifikationsbias*	❑	❑	❑
6. Var referenstestet oberoende av indextestet (dvs indextestet ingick inte som en del av referenstestet)? *Undvikande av inkorporationsbias*	❑	❑	❑
7. Tolkades resultaten från referenstestet utan kännedom om resultaten från indextestet? (Indextestresultat blindade) *Undvikande av informationsbias*	❑	❑	❑
8. Tolkades resultaten från indextestet utan kännedom om resultaten från referenstestet? (Referenstestresultat blindade) *Undvikande av informationsbias*	❑	❑	❑
9. Fanns samma kliniska data tillgängliga då testresultaten tolkades som skulle vara tillgängliga då testen används i praktiken? (Relevant klinisk information)	❑	❑	❑
10. Rapporterades ej tolkningsbara/intermediära testresultat?	❑	❑	❑
11. Förklarades bortfall av patienter från studien?	❑	❑	❑

Förklaring/kommentarer till enskilda kriterier och instruktioner till hur enskilda frågor i mallen ska bedömas och kodas[1]

1. Var sammansättningen av patientgruppen (spektrum) representativ för de patienter som kommer att få testet i praktiken?

Det finns två aspekter på frågan:

- Var den *rekryterade patientgruppen rätt* för att besvara den aktuella frågeställningen?
- Var *metoden för att rekrytera patienter* adekvat så att ett representativt urval erhölls?

Rätt patientgrupp. Det är viktigt att de patienter som ingår i undersökningen har en adekvat sammansättning (patientspektrum), eftersom skillnader i demografiska och kliniska karakteristika mellan populationer kan ge varierande resultat för diagnostisk tillförlitlighet. Om sammansättningen av de patienter som ingått i studien inte motsvarar dem som kommer att bli föremål för testet i klinisk praxis, saknar resultaten relevans. Observera att en studie kan ha god intern validitet men sakna relevans för den aktuella frågeställningen. ”Spektrum” avser inte bara allvarlighetsgraden hos det underliggande sökta tillståndet, utan också hur patienterna valts ut avseende demografiska förhållanden, differentialdiagnos och samsjuklighet. För att bedöma om sammansättningen av patienterna är relevant, kan klinisk information om patienterna, t ex symtom och eventuella föregående tester också vara väsentlig. Sammanfattningsvis är det alltså viktigt att det finns en tydlig beskrivning av den undersökta populationen och tydliga kriterier för inklusion respektive exklusion av patienter. Uppgifter om detta bör man kunna få via rapporterade inklusions- och exklusionskriterier och/eller tabeller avseende patientkarakteristika.

Teoretiskt är sensitivitet och specificitet oberoende av sjukdomsprevalensen i populationen. I praktiken påverkas dock både sensitivitet och specificitet av patientkarakteristika, dvs det patientspektrum som testet appliceras på. Det innebär att sensitivitet och specificitet för ett test på patienter som remitterats till en specialistklinik inte kan förväntas vara desamma som för patienter i allmänpraxis. Den förra utgör en selekterad population, ofta med symtom på (ännu odiagnostiserad) sjukdom, medan den senare primärt är mer oselekterad och med större andel patienter utan sjukdom. Sensitivitet och specificitet påverkas därmed också av sjukdomsprevalensen, som är lägre i en oselekterad population [3]. I regel är sensitiviteten lägre och specificiteten högre i en oselekterad population [3–5].

[1] Efter Whiting [1] och Reitsma [2].

Metod för att rekrytera patienter. Det är också viktigt att patienterna rekryterades på ett adekvat sätt. Den bästa studiedesignen för att få ett representativt urval är ett prospektivt och konsekutivt urval av patienter som uppfyller valda kriterier för inklusion. Så kallade fall–kontrollstudier där man jämför en grupp sjuka patienter med en grupp friska kontrollindivider övervärderar den diagnostiska tillförlitligheten.

Relevans. Kriteriet är alltid relevant och ska alltid ingå i kvalitetsbedömningen.

Specificera. Granskarna bör i möjligaste mån på förhand specificera vilka patientgrupper som är relevanta utifrån den aktuella frågeställningen (inklusionskriterier). Notera vilka faktorer som kan påverka den diagnostiska tillförlitligheten såsom den kliniska situationen (t ex primärvård, specialistvård, sjukhusvård), allvarlighet hos sjukdomen, sjukdomsprevalens och eventuella tester som föregår det aktuella testet. Om man accepterar en mindre andel ej relevanta patienter, bör storleken på den andelen anges. Här bör också anges om man accepterar studier som innehåller en kontrollgrupp med friska individer.

Hur resultaten ska bedömas. Bedömningen ska baseras på både karakteristika hos de inkluderade patienterna och på metoden som använts för att rekrytera patienter. Frågan besvaras med ”Ja” om man anser att den sammansättning (spektrum) av patienter som ingår i studien är representativ för dem som ska få testet i klinisk praxis, och metoden för rekrytering är adekvat. Om den redovisade populationen/metod för rekrytering inte uppfyller givna villkor bör svaret vara ”Nej”. Studier som rekryterar en grupp friska kontrollpersoner och en grupp där man vet att patienterna har det sökta tillståndet ska under nästan alla förhållanden bedömas med ”Nej”. Om information om patientspektrum och/eller rekryteringsmetod är otillräcklig, besvaras frågan med ”Oklart”.

2. Är det troligt att referenstestet (referensstandard, ”gold standard”) korrekt klassificerar det sökta tillståndet?

Referenstestet används för att bestämma närvaro eller frånvaro av det sökta tillståndet. Skattningar av indextestets diagnostiska förmåga bygger på antagandet att det jämförts med en referenstest som är 100 procent sensitivt och 100 procent specifikt. Om man finner skillnader i resultat mellan indextest och referenstest drar man slutsatsen att indextestet har brister. Valet av referenstest är således av stor betydelse. Tyvärr är perfekta referenstest ovanliga. Ett imperfekt referenstest kan ge upphov till bias avseende den diagnostiska tillförlitligheten hos indextestet. Man bör därför i regel begränsa sin inklusion till studier som baseras på en eller flera acceptabla referenstester.

Om det finns allvarliga betänkligheter, t ex att ett indextest kanske är bättre än tillgängligt referenstest, är sedvanlig beräkning av diagnostisk tillförlitlighet inte längre tillämplig. Under sådana förhållanden bör beräkning av diagnostisk tillförlitlighet inte

göras utan att först noga överväga om det finns alternativa metodologiska metoder som är bättre [5,6].

Relevans. Kriteriet är alltid relevant och ska alltid ingå i kvalitetsbedömningen.

Specificera. Vad som är acceptabel referensstandard måste definieras (inklusionskriterier). Inom vissa områden är referensstandarden bestämd genom konsensus. Ibland används en blandning av referensstandarder, och man kan då behöva överväga om alla är acceptabla.

Hur resultaten ska bedömas. Det är inte alltid självklart hur man ska bedöma referenstestets validitet. I regel behövs klinisk erfarenhet i ämnet för att bedöma om ett test eller en kombination av tester utgör en adekvat referensstandard. Om man bedömer att referenstestet är acceptabelt är svaret ”Ja”. Om man bedömer att referenstestet sannolikt inte korrekt klassificerar det sökta tillståndet bör man svara ”Nej”. Om det saknas tillräcklig information för att bedöma detta, svarar man ”Oklart”.

3. Var tidsintervallet mellan referenstest och indextest så kort att det studerade tillståndet inte kunnat förändras mellan de båda testen?

Bäst är om index- och referenstest genomförs samtidigt på samma patienter. Om så inte var fallet, kan fel uppstå genom att patienterna antingen försämrats eller förbättrats (spontant eller genom behandling) under tidsperioden mellan de båda testen. Betydelsen av tidsintervallets längd varierar beroende på det studerade tillståndet. En fördröjning på några dagar är t ex sannolikt inget problem för kroniska tillstånd, men kan vara oacceptabelt vid akuta infektioner.

Relevans. Kriteriet är relevant i de flesta situationer.

Specificera. Bestäm vad som ska betraktas som acceptabelt intervall mellan indextest och referenstest. Bestäm också om det är acceptabelt att en viss andel (ange storlek) av patienterna ligger utanför intervallet.

Hur resultaten ska bedömas. Det gäller att ta ställning till hur stor risken är för felklassificering. Det är vanligt att tidsintervallet mellan testen varierar mellan ingående patienter. Man bör då utgå från det längsta tidsintervall som förekommit mellan index- och referenstest. Om detta bedöms vara tillräckligt kort bedöms kriteriet vara uppfyllt (”Ja”). Om inte är svaret ”Nej”. Om risken för felklassificering inte går att bedöma, är svaret ”Oklart”.

4. Genomgick samtliga patienter eller ett slumpmässigt urval av patienter det avsedda referenstestet?

På engelska använder man bl a termerna ”partial verification bias”, ”work-up bias”, ”primary selection bias” eller ”sequential ordering bias” för att beteckna situationer där inte alla som genomgick indextestet också genomgick referenstestet. Om resultat från indextestet påverkar beslutet att genomföra referenstestet riskerar man att få snedvridna resultat.

Om patienterna randomiseras till att genomgå/inte genomgå referenstestet påverkas inte resultaten. Det händer dock ofta att urvalet av patienter som genomgår referenstest inte sker slumpmässigt.

Relevans. Partiell verifikationsbias förekommer generellt sett bara i prospektiva kohortstudier, där patienterna genomgår indextestet före referenstestet. I situationer där referenstestet bedöms före indextestet får möjligheten till bias bedömas utifrån den aktuella situationen.

Specificera. Eventuellt kan det vara relevant att bestämma hur stor andel som inte verifieras som kan accepteras.

Hur resultaten ska bedömas. Om det står klart att samtliga patienter eller ett slumpmässigt urval av dem som fått indextestet även verifierades enligt referenstestet, är svaret ”Ja”. Om vissa av de patienter som fick indextestet inte fick sitt sanna tillstånd verifierat och urvalet av patienter som fick referenstestet inte var slumpmässigt, är svaret ”Nej”. Om uppgift saknas kodas ”Oklart”.

5. Användes ett och samma referenstest oberoende av vilket resultat som erhölls på indextestet?

Differentiell verifikation uppstår när patienternas indextester valideras mot olika referenstester. Om dessa referenstester definierar det sökta tillståndet på olika sätt, finns risk för differentiell verifikationsbias. Detta inträffar ofta när patienter som får positivt resultat på indextestet blir föremål för ett mer avancerat, inte sällan invasivt referenstest, jämfört med dem som har ett negativt resultat på indextestet. En sådan situation inträffar t ex när det bedöms som oetiskt att använda ett invasivt referenstest hos individer som fått ett negativt resultat på indextestet. Om negativa testresultat hos ett indextest verifieras med ett mindre korrekt referenstest, kommer detta att påverka den diagnostiska tillförlitligheten. En extrem form av differentiell verifikation är när en del av de negativa indextestresultaten inte verifieras alls. Detta leder till övervärdering av både sensitivitet och specificitet.

Differentiell verifikation kan också uppstå när olika centra använder olika referenstest.

Empiriska studier har visat att differentiell verifikation är en viktig källa till bias [7,8]. För att uppskatta risken för allvarlig bias, är det viktigt att förstå varför olika individer verifierades med olika referenstest och skillnaden i kvalitet mellan de olika referenstesterna. Om valet är relaterat till resultat av indextestet eller till sannolikheten för sjukdom (eller tillståndet ifråga), är bias en reell möjlighet.

Relevans. Risk för differentiell verifikationsbias föreligger alltid i studier av diagnostiska test.

Specificera. I regel behövs inga detaljer.

Hur resultaten ska bedömas. Om referenstestet genomgående varit detsamma kodas ”Ja”. Om patienterna genomgått alternativa referenstest blir svaret på frågan ”Nej”. När uppgift saknas kodas ”Oklart”.

6. Var referenstestet oberoende av indextestet (dvs indextestet ingick inte som en del av referenstestet)?

Ibland bestäms referensstandarden med hjälp av flera komponenter eller baseras på information som samlats in under en längre period (t ex en diagnos hos en patient som skrivs ut från sjukhus). När resultatet från indextestet också är inkorporerat i underlaget för att fastställa diagnos (referensstandard), kommer detta att överskatta värdet av indextestet (engelska: ”incorporation bias”). Ett exempel är en studie där man undersökte den diagnostiska tillförlitligheten hos MRI (”magnetic resonance imaging”) för att diagnostisera multipel skleros. Referensstandarden, den slutliga diagnosen, baserades på all tillgänglig information inkluderande resultat från MRI, analys av cerebrospinalvätska (CFS) och klinisk uppföljning av patienten.

Relevans. Kriteriet är endast av betydelse när referenstestet utgörs av flera komponenter. Då är det viktigt att en fullständig definition lämnas av hur tillståndet fastställts och med vilka test detta skett. I studier där referenstestet utgörs av en enda undersökning saknar kriteriet relevans och ska antingen kodas ”Ja” eller utelämnas från kvalitetsbedömningen.

Specificera. I regel behövs inga detaljer här.

Hur resultaten ska bedömas. Om det förefaller som om indextestet utgjorde en del av referenstestet kodas ”Nej”, annars kodas ”Ja”.

7 och 8. Tolkades resultaten från referenstestet utan kännedom om resultaten från indextestet? Tolkades resultaten från indextestet utan kännedom om resultaten från referenstestet?

Denna fråga motsvarar ”blindning” i behandlingsstudier. Tolkning av testresultat kan påverkas av kännedom om resultatet från det alternativa testet. Detta kallas på engelska för ”test review bias” och kan leda till att värdet av testet överskattas. I vilken mån denna bias påverkar resultaten beror framför allt på graden av subjektivitet i tolkningen. Ju mer utrymme för subjektivitet desto större risk för ”review bias”. Det är därför angeläget att bedöma i vilken mån kännedom om resultatet från det ena testet kunnat påverka tolkningen av det andra testet. Huruvida blindning använts eller inte redovisas inte alltid explicit. Vid några tillfällen, t ex när laboratorietester skickas till ett oberoende laboratorium, kan man anta att testet tolkas oberoende av referenstestet. Konfirmering om blindning från författarna är dock alltid önskvärd.

Relevans. Kriteriet är alltid relevant och ska alltid ingå i kvalitetsbedömningen. I de fall testresultaten är fullständigt objektiva (t ex mätvärden) eller utvärderingen görs på ett oberoende laboratorium är risken för ”review bias” liten.

Specificera. I regel behövs inga detaljer här.

Hur resultaten ska bedömas. Om det tydligt framgår att testresultaten tolkats blint blir svaret på frågan ”Ja”. Om det klart framgår att testresultat tolkades med kännedom om det alternativa testet är svaret ”Nej”. Om det är oklart om blindning tillämpades svarar man ”Oklart”.

9. Fanns samma kliniska data tillgängliga då testresultaten tolkades som skulle vara tillgängliga då testen används i praktiken?

För vissa indextest kan tillgång till anamnestiska/kliniska data (t ex ålder, närvaro och allvarlighet av kliniska symtom eller andra testresultat) påverka resultatet, särskilt om indextestet förutsätter tolkning. Ett exempel är tolkning av bilder, som kan påverkas av kännedom om förekomst, karaktär och lokalisering av symtom. Om sådana kliniska data finns tillgängliga när indextestet bedöms, bör de också finnas tillgängliga i klinisk praxis (och vice versa). Det kan vara svårt att separera det diagnostiska värdet av existerande klinisk information innan testet görs från det adderade värdet av indextestet. Hur detta ska hanteras får bedömas utifrån den aktuella frågeställningen. Tillgång till klinisk information hos den som bedömer ett test (gäller i huvudsak röntgenbilder) ökar den diagnostiska sensitiviteten, medan specificiteten inte försämras [3,9].

Om testet avses ersätta andra kliniska test, bör resultat från dessa inte finnas tillgängliga vid tolkningen.

Relevans. Om indextestet avser objektiva mätningar (t ex biokemiska analyser), som inte förändras pga extern information, är risken för bias liten, och kriteriet saknar relevans.

Specificera. Ange vilka kliniska data som normalt sett är tillgängliga i klinisk praxis när testet görs och tolkas, eller alternativt att ingen information vanligtvis är tillgänglig.

Hur resultaten ska bedömas. Om kliniska data som vanligen finns tillgängliga när testresultaten ska tolkas och liknande data fanns tillgängliga i studien svaras ”Ja”. Om tillgängliga kliniska data har undanhållits, eller om mer information än som vanligen finns tillgänglig är svaret ”Nej”. Om information om tillgängliga kliniska data inte rapporteras är svaret ”Oklart”.

10. Rapporterades ej tolkningsbara/intermediära testresultat?

Ett diagnostiskt test kan vara ofullständigt eller ge resultat som inte kan tolkas av olika skäl. Detta kan förekomma i varierande omfattning beroende på testets egenskaper. Problem av denna karaktär rapporteras sällan i studier av diagnostiska test. De utesluts helt enkelt från analysen. Detta kan leda till felaktiga bedömningar. I vilken mån detta inträffar beror på korrelationen mellan ej tolkningsbara resultat och förekomst av positiva resultat enligt referenstestet. Oavsett orsakerna till varför vissa resultat inte kunde tolkas, är det viktigt att förekomsten av sådana problem redovisas i kvantitativ form.

Relevans. Kriteriet är alltid relevant och ska alltid ingå i kvalitetsbedömningen.

Specificera. I regel behövs inga detaljer här.

Hur resultaten ska bedömas. Om man bedömer att det finns ett bortfall av ej tolkningsbara resultat, och att bortfallet kan vara korrelerat med sanna positiva resultat enligt referenstestet, blir svaret på frågan ”Nej”.

11. Förklarades bortfall av patienter från studien?

Bortfall av patienter kan förekomma för ett eller båda testerna. Om bortfallet skiljer sig systematiskt mellan dem som finns kvar och dem som förloras (oavsett anledning), kan utfallet av testresultaten bli snedvridna.

Relevans. Kriteriet är alltid relevant och ska alltid ingå i kvalitetsbedömningen.

Specificera. I regel behövs inga detaljer här.

Hur resultaten ska bedömas. Om det framgår hur stort bortfallet varit för de använda testen, t ex i form av ett flödesdiagram, svaras ”Ja”. Om det framgår att några av de individer som primärt inkluderades inte genomgick både indextest och referenstest, och dessa individer inte har beaktats i analysen, svaras ”Nej”. Om det är oklart hur bortfallet hanterades, svaras ”Oklart”.

Referenser

1. 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.

2. Reitsma JB, Rutjes AWS, Whiting P, Vlassov VV, Leeflang MMG, Deeks JJ. Chapter 9: Assessing methodological quality. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.0. The Cochrane Collaboration, 2009. http://srdta.cochrane.org/

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. Leeflang MM, Bossuyt PM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. J Clin Epidemiol 2009;62:5-12.

5. Knottnerus JA. Diagnostic prediction rules: principles, requirements and pitfalls. Prim Care 1995;22:341-363.

6. Glasziou P, Irwig L, Deeks JJ. When should a new test become the current reference standard? Ann Intern Med 2008;149:816-22.

7. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, Bossuyt PM. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061-6.

8. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006;174:469-76.

9. Loy CT, Irwig L. Accuracy of diagnostic tests read with and without clinical information: a systematic review. JAMA 2004;292:1602-9.

Bilaga 5. **Mall för kvalitetsgranskning av studier med kvalitativ forskningsmetodik – patientupplevelser**

VERSION 2012:1.4

SBU:s granskningsmall bygger på tidigare publicerat material [1,2], men har bearbetats och kompletterats för att passa SBU:s arbete.

Författare: _____ År: _____ Artikelnummer: _____

Total bedömning av studiekvalitet:		
Hög ❑	Medelhög ❑	Låg ❑

Anvisningar:

- Alternativet ”oklart” används när uppgiften inte går att få fram från texten.
- Alternativet ”ej tillämpligt” väljs när frågan inte är relevant.

	Ja	**Nej**	**Oklart**	**Ej tillämpl**
1. Syfte				
a) Utgår studien från en väldefinierad problemformulering/frågeställning?	❑	❑	❑	❑
Kommentarer (syfte, problemformulering, frågeställning etc):				
2. Urval				
a) Är urvalet relevant?	❑	❑	❑	❑
b) Är urvalsförfarandet tydligt beskrivet?	❑	❑	❑	❑
c) Är kontexten tydligt beskriven?	❑	❑	❑	❑
d) Finns relevant etiskt resonemang?	❑	❑	❑	❑
e) Är relationen forskare/urval tydligt beskriven?	❑	❑	❑	❑
Kommentarer (urval, patientkarakteristika, kontext etc):				

	Ja	Nej	Oklart	Ej tillämpl
3. Datainsamling				
a) Är datainsamlingen tydligt beskriven?	❑	❑	❑	❑
b) Är datainsamlingen relevant?	❑	❑	❑	❑
c) Råder datamättnad?	❑	❑	❑	❑
d) Har forskaren hanterat sin egen förförståelse i relation till datainsamlingen?	❑	❑	❑	❑
Kommentarer (datainsamling, datamättnad etc):				
4. Analys				
a) Är analysen tydligt beskriven?	❑	❑	❑	❑
b) Är analysförfarandet relevant i relation till datainsamlingsmetoden?	❑	❑	❑	❑
c) Råder analysmättnad?	❑	❑	❑	❑
d) Har forskaren hanterat sin egen förförståelse i relation till analysen?	❑	❑	❑	❑
Kommentarer (analys, analysmättnad etc):				
5. Resultat				
a) Är resultatet logiskt?	❑	❑	❑	❑
b) Är resultatet begripligt?	❑	❑	❑	❑
c) Är resultatet tydligt beskrivet?	❑	❑	❑	❑
d) Redovisas resultatet i förhållande till en teoretisk referensram?	❑	❑	❑	❑
e) Genereras hypotes/teori/modell?	❑	❑	❑	❑
f) Är resultatet överförbart till ett liknande sammanhang (kontext)?	❑	❑	❑	❑
g) Är resultatet överförbart till ett annat sammanhang (kontext)?	❑	❑	❑	❑
Kommentarer (resultatens tydlighet, tillräcklighet etc):				

Kommentarer till mallen för kvalitetsgranskning av studier med kvalitativ forskningsmetodik – patientupplevelser

1. Syfte

Fundera över:

- vad målsättningen med studien var
- varför det är viktigt
- relevansen
- om kvalitativ metodik är lämplig för att utforska problem-området/svara på frågeställningen.

2. Urval

Fundera över:

- om forskaren redovisat bakgrund till vald urvalsmetod
- om forskaren redovisat hur deltagarna valdes ut
- om forskaren redovisat varför de valda deltagarna valdes ut
- om forskaren redovisat hur många deltagare som valdes ut
- om forskaren redogjort för om någon inte valde att delta och i så fall varför
- om forskaren lyfter fram etiska resonemang som sträcker sig längre än ”informed consent” och ”ethical approval”
- om forskaren beskrivit relationen mellan forskare och informant och hur denna skulle kunna påverka datainsamlingen, exempelvis tacksamhetsskuld, beroendeförhållanden etc.

3. Datainsamling

Fundera över:

- om ”settingen” för datainsamlingen var berättigad
- om det framgår på vilket sätt datainsamlingen utfördes (t ex djup-intervju, semistrukturerad intervju, fokusgrupp, observationer etc)
- om forskaren har motiverat vald datainsamlingsmetod
- om det explicit framgår hur vald datainsamlingsmetod utfördes (t ex vem intervjuade, hur länge, användes intervjuguide, var utfördes intervjun, hur många observationer etc)
- om metoden modifierades under studiens gång (om så är fallet, framgår det hur och varför detta skedde)
- om insamlat datamaterial är tydliga (t ex video- eller ljudinspelningar, anteckningar etc)

- om forskaren resonerar kring om man nått mättnad, dvs när mer datainsamling inte ger mer ny data (inte alltid tillämpbart)
- om det är tillämpbart att föra ett mättnadsresonemang, fundera på om det är rimligt, dvs faktiskt validerat på goda grunder.

4. Analys

Fundera över:

- om analysprocessen är beskriven i detalj
- om analysförfarandet är i linje med den teoretiska ansats som eventuellt låg till grund för datainsamlingen
- om analysen är tematisk, framgår det hur man kommit fram till dessa teman?
- om tabeller har använts för att tydliggöra analysprocessen
- om forskaren kritiskt har resonerat kring sin egen roll, potentiell bias eller inflytande under analysprocessen
- om analysmättnad råder (kan man hitta fler teman baserat på redovisade citat?).

5. Resultat

Fundera över:

- om resultaten/fynden diskuteras i relation till syftet eller frågeställningen
- om ett adekvat resonemang förs kring resultaten eller om resultaten bara är citat/dataredovisning
- om resultaten redovisas på ett tydligt sätt (t ex är det lätt att se vad som är citat/data och vad som är forskarens eget inlägg)
- om resultatredovisningen återkopplas till den teoretiska ansats som eventuellt låg till grund för datainsamling och analys
- om tillräckligt med data redovisas för att underbygga resultaten
- i vilken utsträckning motstridiga data har beaktats och framhålls
- om forskaren kritiskt har resonerat kring dess egen roll, potentiell bias eller inflytande under analysprocessen
- om forskaren för ett resonemang kring resultatens överförbarhet eller andra användningsområden för resultaten.

Referenser

1. Bahtsevani C. In search of evidence-based practices: exploring factors influencing evidence-based practice and implementation of clinical practice guidelines. Malmö: Malmö högskola; 2008.
2. Willman A, Stoltz P, Bahtsevani C. Evidensbaserad omvårdnad. En bro mellan forskning och klinisk verksamhet. Studentlitteratur; 2006.

Bilaga 6. **Mall för kvalitetsgranskning av systematiska översikter enligt AMSTAR** [1,2]

VERSION 2012:1

AMSTAR ger en beskrivning av hur författarna har genomfört en systematisk översikt och om översikten uppfyller grundläggande kvalitetskrav.

Författare: _____ År: _____ Artikelnummer: _____

	Ja	Nej	Kan inte svara	Ej tillämpl
1. Redovisas en förutbestämd metod för genomförandet? Forskningsfrågan och inklusionskriterierna ska vara fastställda innan översikten genomförs.	❑	❑	❑	❑
2. Gjordes studieurval och dataextraktion av två oberoende granskare? Minst två oberoende granskare ska ha utfört dataextraktionen, och ett konsensusförfarande bör vara definierat för att lösa oenigheter.	❑	❑	❑	❑
3. Var litteratursökningen av tillfredsställande omfattning? Sökningen bör göras i minst två elektroniska databaser. Översikten ska ange de årtal och databaser som ingår (t ex Central, Embase och Medline). Ämnesord (”keywords”) och/eller MeSH-termer ska anges och i tillämpliga fall sökstrategin. Alla sökningar bör kompletteras med genomgång av översiktsartiklar, läroböcker, aktuella innehållsförteckningar, ämnesspecifika databaser och register eller rådfrågning av experter, samt av referenslistorna i de framtagna studierna.	❑	❑	❑	❑
4. Användes studiernas publikationsform som ett inklusions-/exklusionskriterium? Författarna bör ange om alla typer av publikationer omfattades av litteratursökningen. Om litteratur har exkluderats pga publikationsform (t ex ”grå litteratur”) eller pga språk, etc ska detta anges.	❑	❑	❑	❑
5. Finns förteckningar över inkluderade och exkluderade studier? En förteckning över medtagna respektive uteslutna studier bör finnas i rapporten.	❑	❑	❑	❑

	Ja	Nej	Kan inte svara	Ej tillämpl
6. Har de inkluderade studiernas karakteristika och resultat redovisats? Kända faktorer hos deltagarna i de utvärderade studierna (”patient characteristics”), såsom ålder, etnicitet, kön, relevanta socioekonomiska data, sjukdomstillstånd, varaktighet, svårighetsgrad och andra sjukdomar, bör anges i rapporten. Uppgifter om deltagarna, åtgärd/behandling och utfall i studierna bör presenteras i sammanfattad form, t ex i en tabell.	❑	❑	❑	❑
7. Har den vetenskapliga kvaliteten hos de ingående studierna utvärderats och dokumenterats? Förutbestämda metoder för kvalitetsvärderingen ska anges. För effektstudier bör exempelvis framgå om författarna valt att bara ta med randomiserade, dubbelblindade studier med kontrollgrupper som får placebo. För andra studietyper gäller andra ställningstaganden.	❑	❑	❑	❑
8. Har vederbörlig hänsyn tagits till de inkluderade studiernas vetenskapliga kvalitet vid formulering av slutsatserna? Utvärderingen av metodologisk stringens och vetenskaplig kvalitet ska framgå i översiktens analys och dess slutsatser, och tydligt anges vid utformning av rekommendationer.	❑	❑	❑	❑
9. Användes lämpliga metoder för sammanvägning av studiernas resultat? Lämpligheten i att lägga samman resultaten från de olika studierna bör säkerställas genom bedömning av de ingående studiernas homogenitet (dvs Chi-2-test för beräkning av homogenitet, I^2). Om heterogenitet finns bör man använda en modell som tar hänsyn till slumpeffekter (”random effects model”) och/eller överväga om det ur klinisk synpunkt är lämpligt att slå ihop resultaten.	❑	❑	❑	❑
10. Har sannolikheten för publikationsbias* bedömts? En bedömning av publikationsbias bör omfatta en kombination av grafiska hjälpmedel (t ex med ”funnel plot” eller andra tester) och/eller statistiska metoder (t ex Eggers regressionsanalys).	❑	❑	❑	❑
11. Är eventuella intressekonflikter angivna? Eventuella sponsorer och bidragsgivare bör tillkännages både i den systematiska översikten och i de ingående studierna.	❑	❑	❑	❑

* SBU:s kommentar: Publikationsbias leder till snedvriden publikation, t ex att positiva resultat publiceras oftare än negativa resultat.

Referenser

1. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews.BMC Medical Research Methodology 2007,7:10.

2. Shea BJ, Bouter LM, Peterson J, Boers M, Andersson N, Ortiz Z, et al. External validation of a measurement tool to assess systematic reviews (AMSTAR). 2007. PLoS ONE 2:e1350.

Bilaga 7. **Mall för kvalitetsgranskning av empiriska hälsoekonomiska studier**

VERSION 2012:1

SBU:s granskningsmall för empiriska hälsoekonomiska studier bygger på tidigare checklistor [1–3] men har bearbetats och kompletterats för att passa SBU:s arbete.

Vägledning för bedömning av studiens relevans, överförbarhet och kvalitet

Eftersom frågorna i Avsnitt 1 berör studiens relevans för projektet är det för att fortsätta med bedömningen enligt frågorna i Avsnitt 2–4 en förutsättning att alla frågorna i Avsnitt 1 fått ett ja-svar. Avsnitt 2 handlar om studiens överförbarhet när det gäller de ekonomiska resultaten. Studiens kvalitet bedöms i Avsnitt 3 och 4. Endast ett fåtal hälsoekonomiska analyser uppfyller checklistans krav i sin helhet. Det innebär inte att studier som inte motsvarar alla krav skulle vara utan värde, men däremot att man bör vara medveten om bristerna vid tolkning av resultaten. En helhetsbedömning avseende studiens överförbarhet respektive kvalitet görs i nedanstående rutor efter att formuläret har fyllts i.

Författare: _____ År: _____ Artikelnummer: _____

Bedömning av överförbarhet av studiens ekonomiska resultat (Avsnitt 2):			
Hög ❑	Medelhög ❑	Låg ❑	Otillräcklig ❑
Bedömning av studiens kvalitet vad gäller *ekonomiska aspekter* (Avsnitt 3 och 4):			
Hög ❑	Medelhög ❑	Låg ❑	Otillräcklig ❑
Bedömning av studiens kvalitet vad gäller *medicinska data*: (projektets medicinska experter avgör)			
Hög ❑	Medelhög ❑	Låg ❑	Otillräcklig ❑

1. Frågor om studiens relevans (”PICO”) i förhållande till projektets frågeställningar *krav på Ja-svar för inklusion*	**Ja**	**Nej**	**Oklart**	**Ej relevant**
a) Är studerad patientpopulation relevant?	❑	❑	❑	❑
b) Är interventionen relevant?	❑	❑	❑	❑
c) Är jämförelseinterventionen relevant?	❑	❑	❑	❑
d) Är utfallsmåttet relevant?	❑	❑	❑	❑

2. Frågor om överförbarhet av studiens ekonomiska resultat	Ja	Nej	Oklart	Ej relevant
a) Studeras både kostnader och effekter (eller anges lika effekt)?	❑	❑	❑	❑
b) Är sjukvårdsorganisationen relevant för svenska förhållanden?	❑	❑	❑	❑
c) Är kostnaderna som används i studien relevanta för nutida svensk sjukvård?	❑	❑	❑	❑
d) Är studiens resultat överförbart till det sammanhang som frågeställningen gäller? [1]	❑	❑	❑	❑
e) Har studien ett samhällsperspektiv?	❑	❑	❑	❑

3. Granskning av eventuella intressekonflikter	Ja	Nej	Oklart	Ej relevant
a) Föreligger, baserat på författarnas angivna bindningar och jäv, låg risk att studiens resultat har påverkats av intressekonflikter?	❑	❑	❑	❑
b) Föreligger, baserat på uppgifter om studiens finansiering, låg risk att studien har påverkats av en finansiär med ekonomiskt intresse i resultatet?	❑	❑	❑	❑
c) Föreligger låg risk för annan form av intressekonflikt (t ex att författarna har utvecklat interventionen)?	❑	❑	❑	❑

4. Frågor för bedömning av studiens kvalitet vad avser den ekonomiska analysen	Ja	Nej	Oklart	Ej relevant
4.1 Val av analys och redovisning av resultat				
a) Är vald form av ekonomisk analys motiverad med avseende på frågeställningarna?	❑	❑	❑	❑
b) Har en inkrementell analys gjorts av både kostnader och effekter (eller går det att räkna fram)?	❑	❑	❑	❑
c) Har lämpliga statistiska metoder använts?	❑	❑	❑	❑
d) Är slutsatserna berättigade med avseende på presenterade resultat?	❑	❑	❑	❑
e) Är tidsperspektivet tillräckligt långt för att ta hänsyn till alla relevanta skillnader i kostnader och effekter?	❑	❑	❑	❑

4. Fortsättning	Ja	Nej	Oklart	Ej relevant
4.2 Effekter och kostnader				
a) Är skillnaden i effekt mellan alternativen som jämförs statistiskt signifikant?	❑	❑	❑	❑
b) Har studien tagit hänsyn till patientföljsamhet (”compliance”)? [2]	❑	❑	❑	❑
c) Har rapporterade data (kostnader och effekter) ett acceptabelt bortfall? [3]	❑	❑	❑	❑
d) Har alla relevanta effekter identifierats (inklusive biverkningar)?	❑	❑	❑	❑
e) Är effekterna kvantifierade på ett lämpligt sätt?	❑	❑	❑	❑
f) Är effekterna på livskvalitet trovärdigt värderade? [4]	❑	❑	❑	❑
g) Har alla relevanta kostnader identifierats, givet tillämpat perspektiv (inklusive biverkningar)?	❑	❑	❑	❑
h) Har kostnaderna mätts på ett korrekt sätt i fysiska enheter (t ex i antal läkarbesök eller antal vårddagar)?	❑	❑	❑	❑
i) Är kostnaderna trovärdigt värderade?	❑	❑	❑	❑
4.3 Känslighetsanalys				
a) Har känslighetsanalys utförts avseende alla betydelsefulla variabler? [5]	❑	❑	❑	❑
b) Har resultatets osäkerhet undersökts med hjälp av probabilistisk analys?	❑	❑	❑	❑
c) Är utfallet robust för undersökta variabelvärden? [6]	❑	❑	❑	❑
4.4 Diskontering (vid studier längre än 1 år) [7]				
a) Har kostnaderna diskonterats på lämpligt sätt?	❑	❑	❑	❑
b) Har effekterna diskonterats på lämpligt sätt?	❑	❑	❑	❑

Eventuella kommentarer till studien: _____

_____

_____

Mall för kvalitetsgranskning av empiriska hälsoekonomiska studier: förklaringar

1. Är studien utförd i samma sjukvårdssektor (t ex inom primärvård eller specialistvård) som frågeställningen gäller? Stämmer den vård som patienterna får i studien överens med de patienter som frågeställningen gäller?

2. Har studien tagit hänsyn till patientföljsamhet (dvs ”compliance”, eventuellt kompletterat med uppgift om analys enligt ”intention to treat” (ITT) eller ”last observation carried forward” (LOCF))?

3. Bortfallet för data på kostnader och livskvalitet är inte alltid samma som för kliniska data. Ett generellt stort bortfall, skillnader i bortfallstorlek samt framför allt orsaksskillnader till bortfall ökar risken för bias. Det bortfall som bedöms här avser bortfall efter randomisering. Man kan aldrig räkna med att bortfall är slumpmässigt. Problemet minskar om sammansättningen av personer i bortfallet inte skiljer från dem som finns kvar i studien. Nedanstående exempel kan tjäna som grova riktvärden: litet (<10 %), måttligt (10–19 %), stort (20–29 %) mycket stort (≥30 %). Vid bortfall >30 procent bedöms resultatet ofta sakna informationsvärde vilket kan innebära att studien bör exkluderas.

4. Exempelvis: Vilken tariff användes för att ta fram vikter för kvalitetsjusterade levnadsår (QALY-vikter)? Har värderingar med ”willingness-to-pay”-metoder gjorts på ett trovärdigt sätt?

5. Gäller variabler där det råder osäkerhet och som kan förväntas påverka analysen. Om extrapoleringar gjorts utifrån empiriska data kan det vara viktigt att testa olika sätt att extrapolera.

6. Med robust menas att resultatet inte ändras så pass mycket i känslighetsanalysen att slutsatserna om kostnadseffektivitet ändras (gäller både envägs- och probabilistisk känslighetsanalys).

7. Argumenteras för vald metod på ett adekvat sätt? Olika länder har olika rekommendationer. Framtida kostnader ska diskonteras (men räntan kan variera). För effekter finns det argument både för och emot diskontering. I England och Wales (NICE) används en diskonteringsränta på 3,5 procent på både kostnader och effekter. I Nederländerna används istället 4 procent på kostnader och 1,5 procent på effekter. Tandvårds- och läkemedelsförmånsverket (TLV) rekommenderar en diskonteringsränta på 3 procent på både effekter och kostnader men efterfrågar känslighetsanalyser i vilka räntan sätts till 0 och 5 procent.

Referenser

1. Brunetti M, Ruiz F, Lord J, et al. Chapter 10: Grading economic evidence. In: Schemilt I, Mugford M, Vale L, et al, editors. Evidence-based decisions and economics: health care, social welfare, education and criminal justice. Oxford: Wiley-Blackwell, 2010.

2. Drummond MF, Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL. Methods for the economic evaluation of health care programmes, 3rd edition. Oxford: Oxford University Press, 2005.

3. Evers S, Gossen M, de Vet H, van Tulder M, Ament A. Criteria list for assessment of methodological quality of economic evaluations: Consensus on health economic criteria. International Journal of Technology Assessment in Health Care 2005;21(2):240-5.

Bilaga 8. **Mall för kvalitetsgranskning av hälsoekonomiska modellstudier**

VERSION 2012:1

SBU:s granskningsmall för hälsoekonomiska modellstudier bygger på tidigare checklistor [1–4] men har bearbetats och kompletterats bl a med specifika kriterier för bedömning av modellstudier. För bedömning av kvalitet på data som använts i modellen hänvisas till Cooper och medarbetare [5].

Vägledning för bedömning av studiens relevans, överförbarhet och kvalitet

Eftersom frågorna i Avsnitt 1 berör studiens relevans för projektet är det för att fortsätta med bedömningen enligt frågorna i Avsnitt 2–4 en förutsättning att alla frågorna i Avsnitt 1 fått ett ja-svar. Avsnitt 2 handlar om studiens överförbarhet och relevans när det gäller de ekonomiska resultaten. Studiens kvalitet bedöms i Avsnitt 3 och 4. Endast ett fåtal hälsoekonomiska analyser uppfyller checklistans krav i sin helhet. Det innebär inte att studier som inte motsvarar alla krav skulle vara utan värde, men däremot att man bör vara medveten om bristerna vid tolkning av resultaten. En helhetsbedömning avseende studiens överförbarhet respektive kvalitet görs i nedanstående rutor efter att formuläret har fyllts i.

Författare: _____ År: _____ Artikelnummer: _____

Bedömning av överförbarhet av studiens ekonomiska resultat (Avsnitt 2):			
Hög ❑	Medelhög ❑	Låg ❑	Otillräcklig ❑
Bedömning av studiens kvalitet vad gäller *ekonomiska aspekter* (Avsnitt 3 och 4):			
Hög ❑	Medelhög ❑	Låg ❑	Otillräcklig ❑
Bedömning av studiens kvalitet vad gäller *medicinska data*: (projektets medicinska experter avgör)			
Hög ❑	Medelhög ❑	Låg ❑	Otillräcklig ❑

1. Frågor om studiens relevans (”PICO”) i förhållande till projektets frågeställningar *krav på Ja-svar för inklusion*	**Ja**	**Nej**	**Oklart**	**Ej relevant**
a) Är studerad patientpopulation relevant?	❑	❑	❑	❑
b) Är interventionen relevant?	❑	❑	❑	❑
c) Är jämförelseinterventionen relevant?	❑	❑	❑	❑
d) Är utfallsmåttet relevant?	❑	❑	❑	❑

2. Frågor om överförbarhet av studiens ekonomiska resultat	Ja	Nej	Oklart	Ej relevant
a) Studeras både kostnader och effekter (eller anges lika effekt)?	❑	❑	❑	❑
b) Är sjukvårdsorganisationen relevant för svenska förhållanden?	❑	❑	❑	❑
c) Är kostnaderna som används i studien relevanta för nutida svensk sjukvård?	❑	❑	❑	❑
d) Är studiens resultat överförbart till det sammanhang som frågeställningen gäller? [1]	❑	❑	❑	❑
e) Har studien ett samhällsperspektiv?	❑	❑	❑	❑

3. Granskning av eventuella intressekonflikter	Ja	Nej	Oklart	Ej relevant
a) Föreligger, baserat på författarnas angivna bindningar och jäv, låg risk att studiens resultat har påverkats av intressekonflikter?	❑	❑	❑	❑
b) Föreligger, baserat på uppgifter om studiens finansiering, låg risk att studien har påverkats av en finansiär med ekonomiskt intresse i resultatet?	❑	❑	❑	❑
c) Föreligger låg risk för annan form av intressekonflikt (t ex att författarna har utvecklat interventionen)?	❑	❑	❑	❑

4. Frågor för bedömning av studiens kvalitet vad avser den ekonomiska analysen	Ja	Nej	Oklart	Ej relevant
4.1 Val av analys				
a) Är vald form av ekonomisk analys motiverad med avseende på frågeställningarna?	❑	❑	❑	❑
4.2 Modellstruktur				
a) Är modellstrukturen lämplig för den specifika frågeställningen och det specifika sjukdomstillståndet?	❑	❑	❑	❑
b) Är modellen och eventuella antaganden som gjorts transparenta?	❑	❑	❑	❑
c) Är modellen testad för extern validitet? [2]	❑	❑	❑	❑
d) Är vald tidshorisont tillräckligt lång för att ta hänsyn till alla relevanta skillnader i kostnader och effekter?	❑	❑	❑	❑

4. Fortsättning	Ja	Nej	Oklart	Ej relevant
e) Är vald tidshorisont rimlig i relation till empiriska data?	❑	❑	❑	❑
f) Markov: Är tidscyklernas längd motiverad med avseende på frågeställningen?	❑	❑	❑	❑
4.3 Effekter och kostnader				
a) Är skillnaden i effekt som ligger till grund för modellanalysen statistiskt signifikant?	❑	❑	❑	❑
b) Har studien tagit hänsyn till patientföljsamhet (”compliance”)? [3]	❑	❑	❑	❑
c) Har alla relevanta effekter identifierats (inklusive biverkningar)?	❑	❑	❑	❑
d) Är använda effektdata från bästa möjliga källa? [4]	❑	❑	❑	❑
e) Har alla relevanta kostnader identifierats, givet tillämpat perspektiv (inklusive biverkningar)?	❑	❑	❑	❑
f) Är använda data på förbrukning av resurser (t ex läkarbesök, vårddagar) från bästa möjliga källa?	❑	❑	❑	❑
g) Är uppgifterna om enhetskostnader från bästa möjliga källa?	❑	❑	❑	❑
4.4 Tolkning av resultat				
a) Har inkrementell analys gjorts av både kostnader och effekter (eller går det att räkna fram)?	❑	❑	❑	❑
b) Har lämpliga statistiska metoder använts?	❑	❑	❑	❑
c) Är slutsatserna berättigade med avseende på presenterade resultat?	❑	❑	❑	❑
4.5 Känslighetsanalys				
a) Har känslighetsanalys utförts avseende alla betydelsefulla variabler? [5]	❑	❑	❑	❑
b) Har resultatets osäkerhet undersökts med hjälp av probabilistisk analys?	❑	❑	❑	❑
c) Är utfallet robust för undersökta variabelvärden? [6]	❑	❑	❑	❑
4.6 Diskontering (vid studier längre än 1 år) [7]				
a) Har kostnaderna diskonterats på lämpligt sätt?	❑	❑	❑	❑
b) Har effekterna diskonterats på lämpligt sätt?	❑	❑	❑	❑

Eventuella kommentarer till studien: _____

Mall för kvalitetsgranskning av hälsoekonomiska modellstudier: förklaringar

1. Är studien utförd i samma sjukvårdssektor (t ex inom primärvård eller specialistvård) som frågeställningen gäller? Stämmer den vård som patienterna får i studien överens med de patienter som frågeställningen gäller?

2. Extern validitet innebär oftast att modellens kliniska resultat jämförs med resultat från andra modeller eller kliniska studier. Det kan också innebära att man låtit någon extern person granska modellen ingående. För ett ja-svar räcker inte att studiens inkrementella kostnadseffektkvot (ICER) har jämförts med andra studier.

3. Har studien tagit hänsyn till patientföljsamhet (dvs ”compliance”, eventuellt kompletterat med uppgift om analys enligt ”intention to treat” (ITT) eller ”last observation carried forward” (LOCF))?

4. Finns det fler studier eller studier av bättre kvalitet som innehåller effektdata och bör tas med i analysen?

5. Gäller variabler där det råder osäkerhet och som kan förväntas påverka analysen. Om extrapoleringar gjorts utifrån empiriska data kan det vara viktigt att testa olika sätt att extrapolera.

6. Med robust menas att resultatet inte ändras så pass mycket i känslighetsanalysen att slutsatserna om kostnadseffektivitet ändras (gäller både envägs- och probabilistisk känslighetsanalys).

7. Argumenteras för vald metod på ett adekvat sätt? Olika länder har olika rekommendationer. Framtida kostnader ska diskonteras (men räntan kan variera). För effekter finns det argument både för och emot diskontering. I England och Wales (NICE) används en diskonteringsränta på 3,5 procent på både kostnader och effekter. I Nederländerna används istället 4 procent på kostnader och 1,5 procent på effekter. Tandvårds- och läkemedelsförmånsverket (TLV) rekommenderar en diskonteringsränta på 3 procent på både effekter och kostnader men efterfrågar känslighetsanalyser i vilka räntan sätts till 0 och 5 procent.

Referenser

1. Brunetti M, Ruiz F, Lord J, et al. Chapter 10: Grading economic evidence. In: Schemilt I, Mugford M, Vale L, et al, editors. Evidence-based decisions and economics: health care, social welfare, education and criminal justice. Oxford: Wiley-Blackwell, 2010.

2. Drummond MF, Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL. Methods for the economic evaluation of health care programmes, 3rd edition. Oxford: Oxford University Press, 2005.

3. Evers S, Gossen M, de Vet H, van Tulder M, Ament A. Criteria list for assessment of methodological quality of economic evaluations: Consensus on health economic criteria. International Journal of Technology Assessment in Health Care 2005;21(2):240-5.

4. Philips Z, Ginnelly L, Sculpher M, Claxton K, Golder S, Riemsma R, et al. Review of guidelines for good practice in decision-analytic modeling in health technology assessment. Health technology assessment 2004;8(36):1-72.

5. Cooper N, Coyle D, Abrams K, Mugford M, Sutton A. Use of evidence in decision models: an appraisal of health technology assessments in the UK since 1997. Journal of Health Services Research and Policy 2005;10(4):245-50.

Bilaga 9. **Statistiska begrepp i medicinska utvärderingar**

VERSION 2012:1

Bilagan består av två avsnitt. Det första rör de vanligaste måtten och metoderna för att bedöma validitet och tillförlitlighet i olika diagnostiska metoder. Det andra avsnittet diskuterar olika metoder för att redovisa resultat av behandlingsstudier.

Den som vill ha mer djupgående kunskaper hänvisas till läroböcker i ämnet [1–5]. Referenslistan innehåller en del litteraturtips.

Diagnostiska studier

Det finns tre grundmått: sensitivitet, specificitet och sjukdomsprevalensen i den grupp som undersöks och diagnostiseras. Alla andra mått kan beräknas utifrån dessa tre. Alla mått har sina för- och nackdelar. I detta avsnitt beskriver vi närmare begreppen:

- testmetodens sensitivitet och specificitet
- prediktionsvärden
- ”likelihood”-kvoter
- ROC-kurvor
- reliabilitetsmått.

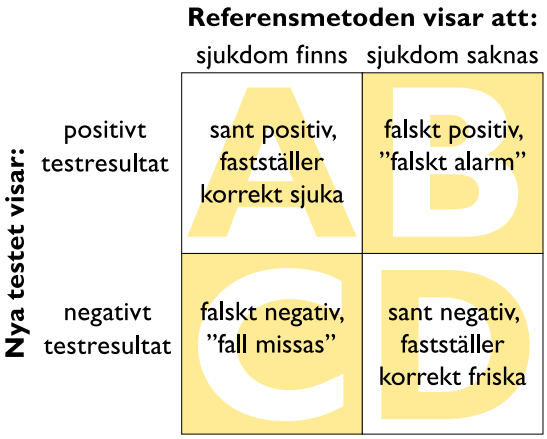
Diagnostik och riskbedömning eller prognos för att förutsäga sjukdomsutvecklingen hör nära samman eftersom diagnostik är en förutsättning för att göra en riskbedömning eller en prognos. Hur träffsäker en diagnostisk metod eller en metod för riskbedömning är, mäts också med samma mått. En bra metod ska vara tillräckligt känslig för att missa så få av dem som är/blir sjuka som möjligt och samtidigt ge så få ”falska alarm” som möjligt, dvs friska/icke riskindivider ska också identifieras med hög träffsäkerhet. Beräkningen förutsätter att man kan jämföra utfallet med någon standard, referensmetod eller det faktiska utfallet. Referensmetoden eller den bästa möjliga referensmetoden som internationellt kallas ”gold standard” varierar.

När man gör en undersökning eller ett test, t ex en kemisk analys, ett cytologiskt prov eller en röntgenundersökning, kan resultaten av tester antingen vara positiva eller negativa.

Positivt test tyder på viss sjukdom. *Negativt test* tyder på hälsa.

Med denna definition kan testresultaten beroende på om patienten verkligen är sjuk eller frisk ge fyra olika utslag:

- Sant positiva = sjuka klassificeras som sjuka (a)
- Sant negativa = friska klassificeras som friska (d)
- Falskt positiva = friska klassificeras som sjuka (b)
- Falskt negativa = sjuka klassificeras som friska (c).



Figur B9.1 Fyrfältstabell med kombinationer av testresultat och sjukdomsförekomst.

Dessa utfall brukar redovisas i en fyrfältstabell (Figur B9.1) med kombinationer av testresultat och sjukdomsförekomst.

Testmetodens sensitivitet och specificitet

Utifrån fyrfältstabellen kan testmetodens tillförlitlighet bedömas med hjälp av två mått, sensitivitet och specificitet. Naturligtvis vill man att testmetoden både ska vara *känslig*, dvs reagera för *alla* med sjukdomen och *specifik*, dvs *bara* reagera för de sjuka. Mått på hur känsligt och specifikt ett test är kallas för sensitivitet respektive specificitet (Faktaruta B9.1).

Faktaruta B9.1 Definitioner och formler för sensitivitet och specificitet.

Sensitivitet = Sannolikheten för positivt testresultat när man har sjukdomen.
Specificitet = Sannolikheten för negativt testresultat när man är frisk.

Sensitivitet = sjuka klassificerade som sjuka/alla sjuka = a/(a+c).
Specificitet = friska klassificerade som friska/alla friska = d/(b+d).

Exempel B9.1 Räkneexempel som illustrerar måtten sensitivitet och specificitet.

	Sjuka (S+)	Friska (S–)	Summa
Positivt test (+)	950	100	1 050
Negativt test (–)	50	900	950
Totalt	1 000	1 000	2 000

Sensitivitet = 950/1 000 = 0,95 = 95 %. Specificitet = 900/1 000 = 0,90 = 90 %.

Sensitivitet och specificitet kan vardera anta värden mellan 0 och 100 procent. Ju närmare 100 procent desto bättre är det diagnostiska/prognostiska testet. Om summan av sensitivitet och specificitet är 2 är testet perfekt, dvs träffsäkerheten (”accuracy”) är 100 procent. Om sensitivitet och specificitet vardera är mindre än 0,5 (träffsäkerhet <50 %), är testet värdelöst, dvs det är inte bättre än slumpen.

När vi ska värdera olika tester ställs vi inför en mängd beslutsproblem. Om ett test har både sensitivitet och specificitet som är högre än ett annat test, så väljer man förstås det första. Oftast får vi dock göra kompromisser och välja antingen hög sensitivitet eller hög specificitet. Det gäller att se vilken typ av feldiagnos som får minst allvarliga konsekvenser.

I en del artiklar har man försökt strukturera de faktorer som har betydelse för vilka test som är lämpliga vid olika tillfällen. Ett exempel på strukturering redovisas i Exempel B9.2.

Exempel B9.2 Strukturering av faktorer som har betydelse för test.

A.	Sjukdomens prevalens (= krav på prevalens)	Förekomst av sjukdom hos den grupp som utsätts för testet
B.	”Skador” av att friska klassificeras som sjuka (= krav på specificitet)	1. Risker med att behandla friska individer 2. Ekonomiska behandlingskostnader 3. Etiska konsekvenser
C.	”Skador” av att sjuka klassificeras som friska (= krav på sensitivitet)	1. Individens risk av att ej bli behandlad 2. Befolkningens risker (smittspridning) 3. Ärftliga risker

Exempel B9.3 Olika krav på sensitivitet och specificitet.

Onödig behandling av syfilis innebär vissa mindre risker och kostnader, men riskerna är på det hela taget små. Kraven på specificitet är därför ganska låga. Däremot är konsekvenserna av obehandlade fall allvarliga både för individen och befolkningen. Det är alltså viktigt att få tag i så många fall som möjligt. Sensitiviteten bör ofta vara hög när det gäller smittsamma sjukdomar.

Lungcancer kan sägas vara en sjukdom med motsatt förhållande. Riskerna vid strålningsbehandling eller operation är stora. Behandlingskostnaderna är också höga. Möjligheterna att bota är för närvarande relativt dåliga och det finns inga risker för befolkningen eller för eventuella arvingar. Specificiteten i testerna bör därför vara hög. Fosterdiagnostik är ett annat exempel där kraven på specificitet måste vara mycket höga.

Det är viktigt att ha ett patientperspektiv när man analyserar diagnostiska metoder. Förutom riskerna med att inte snabbt komma under behandling eller riskerna med att utsättas för felaktig behandling så kan felaktiga eller osäkra besked skapa oro, ängslan eller ångest hos såväl patienter som anhöriga. Denna typ av problem är särskilt accentuerade vid screening där man undersöker en hel grupp där prevalensen av sjukdom är ganska låg. En annan viktig aspekt av patientperspektivet på diagnostik är värdet av ett negativt test. För den enskilde individen är ett negativt testresultat på en cancerundersökning av stor betydelse för att minska oro, och därmed öka välbefinnande. Ett friande negativt test har även en ekonomisk aspekt, se Exempel B9.4. Sammanfattningsvis bör värdet av ett negativt test inte underskattas.

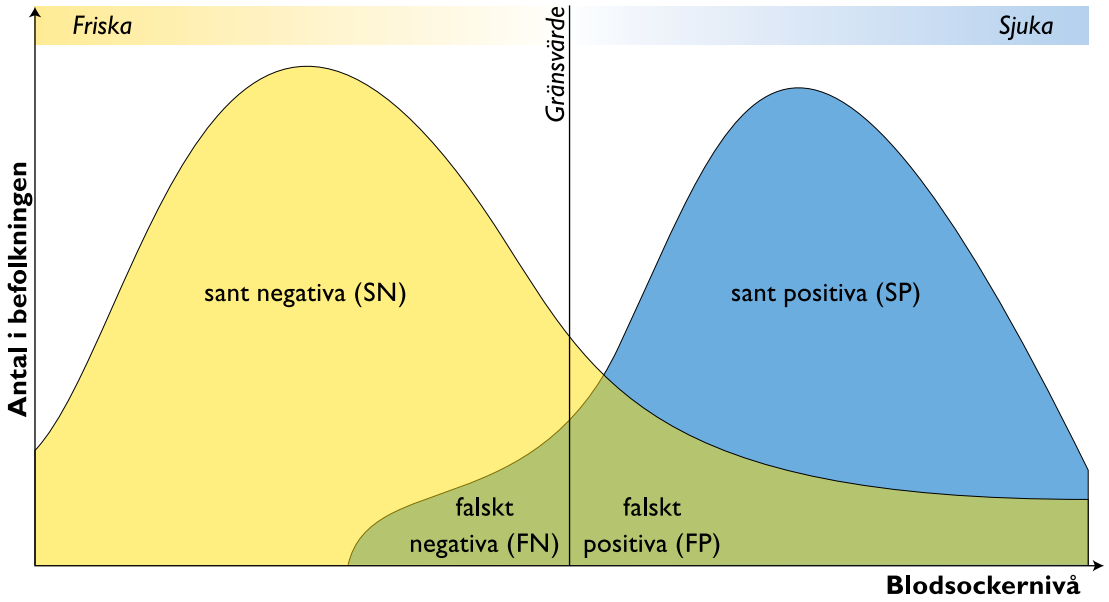
Exempel B9.4 Friande negativa test kan ha fördelar.

En RCT utförd i Danmark [6] visade att 60–70 procent av de patienter som inte undersöktes med gastroskopi (= bästa möjliga referensmetod för utredning av symtom på reflux) kom tillbaka för förnyad undersökning om symtomen återkom, vilket ofta är fallet med dyspepsi.

Resultaten av diagnostiken uttrycks ofta i ett kvantitativt värde, t ex blodtrycks- eller blodsockernivå. Måttligt höga blodtryck är dessutom mer en riskfaktor än en indikation på sjukdom. Gränsvärdet för blodsockernivå avgör vilka som definieras som diabetiker. Gränsen för vad som ska betecknas som sjukt eller friskt är inte självklar. Oftast väljs en gräns, en ”cut-off”-nivå, där man bedömer att risken för sjukdom är stor. Denna något godtyckliga gräns påverkar hur många sant och falskt positiva respektive negativa vi kommer att få (Figur B9.2). Sänks gränsvärdet för blodsockernivån kommer man att få fler falskt positiva och färre falskt negativa. Höjs gränsvärdet blir resultatet det omvända.

I praktiken baseras inte besluten på en testmetod utan oftast på en samlad bedömning av symtom och resultaten av flera diagnostiska tester. Riskerna att göra felslut är begränsade, men det kan ändå vara värt att i varje enskilt fall fundera på om sensitivitet eller specificitet är viktigast.

Sensitivitet och specificitet är två grundläggande mått som visar på testmetodens tillförlitlighet. Ett problem med dessa mått är att de förutsätter att man vet vilka som är sjuka eller friska och att jämförelse- eller referensmetoden förutsätts vara tillförlitlig. Ett annat problem är att det i en klinisk verksamhet inte säger något om sannolikheten att patienten har sjukdomen eller ej, det är bara testresultatet som är känt. Man vill gärna ha mått på möjligheterna att förutsäga (predicera) om patienten är sjuk eller frisk.



FN = Falskt negativa; FP = Falskt positiva; SN = Sant negativa, dvs friska klassificeras som friska; SP = Sant positiva, dvs sjuka klassificeras som sjuka

Figur B9.2 Gränsvärdet för sjukt och friskt påverkar andelen sant/falskt positiva/negativa. Figuren illustrerar fördelningen av en ”frisk” och en ”sjuk” befolkning där det finns en viss överlappning mellan grupperna vad gäller den mätta variabeln. Det lodrätta strecket i mitten är gränsvärdet. Om gränsvärdet förskjuts åt vänster får man fler falskt positiva och förskjuts den åt höger får man fler falskt negativa.

Prediktionsvärden

När man bedömer tillförlitligheten i ett test har man alltså utgått från ett facit, dvs man vet vilka som har en viss sjukdom eller inte. Problemet för läkarna när de ska ställa en diagnos är att de inte vet om patienten har en sjukdom eller inte. Däremot känner de till testresultatet. Det är då mer intressant att veta hur sannolikt det är att patienten har sjukdomen, när testresultatet är positivt. Detta kallar vi *positivt prediktionsvärde* och är en så kallad betingad sannolikhet, dvs sannolikheten för att de som testas har sjukdomen betingat av att testresultatet är positivt.

Ett sätt att beräkna det positiva prediktionsvärdet är att sätta alla sant positiva resultat i relation till alla positiva testresultat, dvs enligt exemplet nedan:

> *Positivt prediktionsvärde (PPV)* = a/a+b = 950 sant positiva/1 050 positiva testresultat = 90,5 %.

På motsvarande sätt är man intresserad av att veta om patienten inte har sjukdomen om testresultatet är negativt. Detta kallas *negativt prediktionsvärde* och är sannolikheten för att de som testas inte har sjukdomen betingat av att testresultatet är negativt, dvs man dividerar sant negativa med alla negativa testresultat. Enligt exemplet nedan får vi:

Negativt prediktionsvärde (NPV) = d/c+d = 900 sant negativa/950 negativa testresultat = 0,947 = 94,7 procent.

Tre faktorer påverkar prediktionsvärdena. Självklart påverkar testmetodens sensitivitet och specificitet möjligheterna att förutsäga om patienten är sjuk eller frisk. Det är inte lika intuitivt självklart att prevalensen i den grupp som undersöks påverkar prediktionsvärdena. Att så dock är fallet kan visas genom att vi ändrar prevalensen i vårt hypotetiska exempel, se Exempel B9.5.

Exempel B9.5 Positiva och negativa prediktionsvärden.

I Exempel B9.1 hade hälften av de undersökta sjukdomen i fråga, dvs en prevalens på 50 procent. Om vi har samma sensitivitet och specificitet, men ändrar sjukdomsprevalensen i den grupp som undersökts till 5 procent blir prediktionsvärdena helt annorlunda.

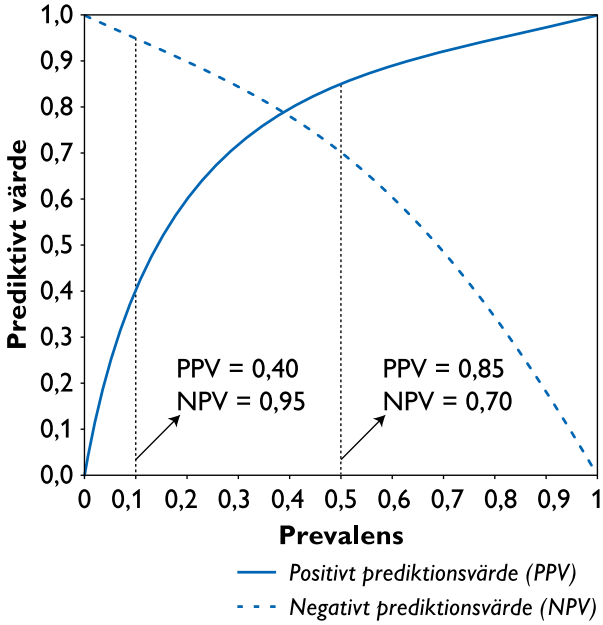
Antal	Sjuka (S+)	Friska (S–)	Summa
Positivt test (+)	95	190	285
Negativt test (–)	5	1 710	1 715
Totalt	100	1 900	2 000

I detta fall med en prevalens på 5 procent blir det positiva prediktionsvärdet (PPV) = 95/285 = 33,3 procent jämfört med 90,5 procent när prevalensen var 50 procent. Det negativa prediktionsvärdet (NPV) blir 99,7 procent (1 710/1 715) jämfört med 94,7 procent när prevalensen var 50 procent. Slutsatsen är att det positiva prediktionsvärdet sjunker om prevalensen är låg medan det negativa prediktionsvärdet blir högre om prevalensen är låg.

Möjligheterna att med en viss testmetod förutsäga om en patient är sjuk eller ej varierar alltså beroende på i vilken situation metoden används. Vid screening där de flesta som testas är friska, dvs en låg prevalens, blir det positiva prediktionsvärdet relativt lågt. Om man testar metoden på en grupp patienter som remitterats pga stark klinisk misstanke om sjukdom kan prevalensen i den gruppen vara mycket hög och man får därigenom ett högt prediktionsvärde. En distriktsläkare träffar kanske bara på sjukdomen i ett fall per 10 000 patienter medan en specialist på universitetssjukhuset får vissa patientkategorier där nästan var femte patient kan ha sjukdomen ifråga. Ett test på en universitetsklinik med selekterade patienter, dvs hög prevalens, fungerar kanske inte lika bra i primärvården där sjukdomsförekomsten är låg. I den kliniska situationen bör man alltså alltid fundera på i vilken situation man befinner sig och bedöma hur prevalensen kan påverka beslutet.

”Likelihood”-kvoter

Det kan vara en fördel att ha ett mått på ett tests prestanda, ett mått som sammanfattar sensitivitet och specificitet och som är oberoende av sjukdomsprevalensen. ”Likelihood”-



Figur B9.3 Diagram som illustrerar hur det positiva och negativa prediktionsvärdet påverkas av sjukdomsprevalensen. I exemplet är sensitiviteten (andelen sanna positiva) 0,6 och specificiteten (andelen sanna negativa) 0,9. Vid en sjukdomsprevalens på 10 procent blir det positiva prediktiva värdet (PPV) 0,40 medan det negativa prediktiva värdet (NPV) är 0,95. Om sjukdomsprevalensen är 50 procent, ökar PPV till 0,85, medan NPV sjunker till 0,70.

kvoter (LHR) är sådana mått som uttrycks i termer av odds. Ett odds är sannolikheten för att en viss händelse ska inträffa dividerad med sannolikheten att den inte ska inträffa. Oddskvoter för ”likelihood”-kvoter redovisas som en sannolikhet för ett visst testresultat om man är sjuk dividerat med sannolikheten för samma testresultat om man är frisk. Eftersom testresultaten antingen kan vara positiva eller negativa kan man få två ”likelihood”-kvoter.

En *positiv ”likelihood”-kvot (LHR+)* beskriver sannolikheten att vara testpositiv om man är sjuk dividerat med sannolikheten att vara testpositiv om man är frisk = *sensitivitet/(1-specificitet).* Det kan också uttryckas som oddset för att ett positivt test kommer från en person med sjukdomen istället för en utan den. Ju högre LHR+-värde desto högre är sannolikheten att personen har sjukdomen i fråga.

En *negativ ”likelihood”-kvot (LHR–)* beskriver sannolikheten att vara testnegativ om man är sjuk dividerat med sannolikheten att vara testnegativ om man är frisk = *(1-sensitivitet)/specificitet.* Det kan också uttryckas som oddset för ett negativt test kommer från en med

sjukdomen istället för en utan den. Ju lägre LHR–-värde desto mindre är sannolikheten att personen har sjukdomen i fråga.

Exempel B9.6 Beräkning av ”likelihood”-kvoter.

Utifrån vårt tidigare fiktiva exempel kan ”likelihood”-kvoterna beräknas enligt följande:

Antal	Sjuka (S+)	Friska (S–)	Summa
Positivt test (+)	950	100	1 050
Negativt test (–)	50	900	950
Totalt	1 000	1 000	2 000

LHR+ = sensitivitet/(1-specificitet) = 0,95/(1–0,90) = 9,5.
Oddset för att ett positivt test kommer från en sjuk person istället för en frisk är 9,5.

LHR– = (1-sensitivitet)/specificitet = (1–0,95)/0,90 = 0,055.
Oddset för att ett negativt test kommer från en sjuk person istället för en frisk är 0,055.

Det är inte helt självklart hur ”likelihood”-kvoter ska tolkas. Grunden är dock att ju högre oddset är för positiva ”likelihood”-kvoter (LHR+) desto bättre är testet på att fastställa sjukdom. När det gäller negativa ”likelihood”-kvoter (LHR–) så är låga odds att föredra eftersom det minskar sannolikheten att personen har sjukdomen. Det finns dock inga klara gränser för vad som är ett stort odds för att personen har sjukdomen. Enkla tumregler har föreslagits, se Faktaruta B9.2.

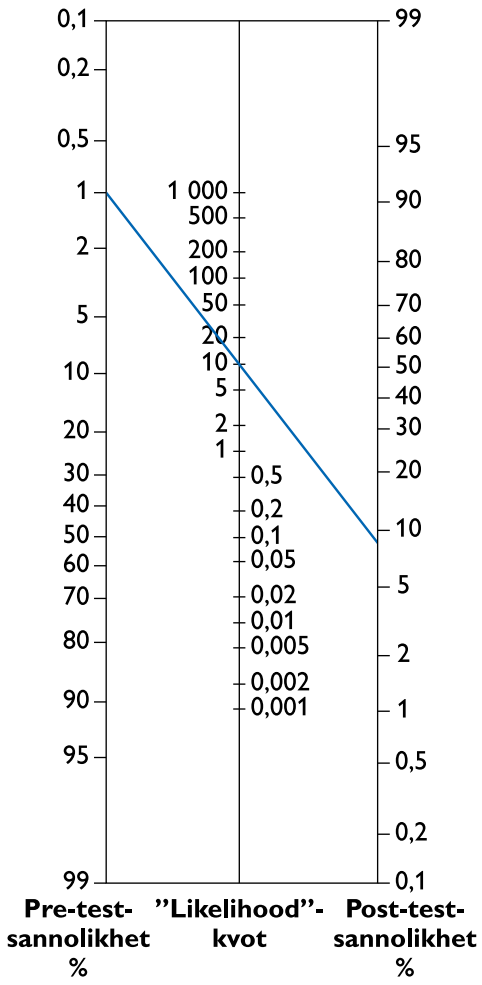
Faktaruta B9.2 Tumregler för ”likelihood”-kvoter.

LHR+	Sannolikhet att personen har sjukdomen
>10	Stor eller mycket stor ökning
5–10	Måttlig ökning
2–5	Liten ökning, kan vara betydelsefull
1–2	Mycket liten ökning, sällan betydelsefull
LHR–	**Sannolikhet att personen har sjukdomen**
<0,1	Mycket stor minskning
0,1–0,2	Måttlig minskning
0,2–0,5	Liten minskning, kan vara betydelsefull
0,5–1	Mycket liten minskning, sällan betydelsefull

LHR+ = Positiv likelihoodkvot; LHR– = Negativ likelihoodkvot

Eftersom diagnostiken oftast bygger på en samlad värdering av symtom, anamnes och flera olika tester kan det vara värdefullt att titta på odds före (pre-test) och odds efter (post-test), dvs före och efter att man fått resultatet av ett test. Matematiskt kan detta uttryckas som produkten av odds före och LHR, dvs *odds efter = odds före x LHR*. Oddset före bildas som prevalensen/(1-prevalensen) och kan antingen baseras på faktiska data om prevalensen i den grupp som studeras eller på subjektiva uppskattningar om risken att patienten i den grupp som studeras har en viss sjukdom.

Oddset kan också redovisas som efter-sannolikhet (”post-test probability”) med formeln *efter-sannolikhet = oddset efter/(oddset efter + 1)*. Observera att oddset efter positivt test är lika med det positiva prediktionsvärdet. För att beräkna oddset efter kan man använda beräkningsprogram eller avläsa det via nomogram (Figur B9.4).



Figur B9.4 Nomogram.

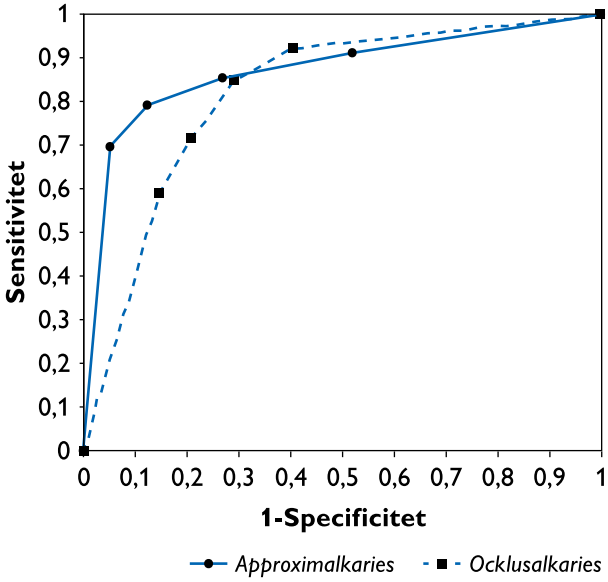
Exempel B9.7 Odds före och efter resultat av test.

I vårt fiktiva exempel med positiva testresultat blir oddset före = 0,5/(1–0,5) = 1. Oddset efter = oddset före x LHR+ = 1x9,5 = 9,5. Efter-sannolikhet = oddset efter/(oddset efter + 1) = 9,5/(9,5 + 1) = 90,5 procent.

ROC (”receiver operating characteristics”)-kurvor

”Receiver operating characteristics” (ROC) är en grafisk redovisning av testmetoders prestanda. Den anger sambandet mellan en testmetods sensitivitet och andel falskt positiva (1-specificitet). Genom att plotta sensitivitet gentemot 1-specificitet för varje ”cut-off” (tröskelvärde) och sedan förbinda punkterna får man en så kallad ROC-kurva. ROC-analys är hämtad från forskning om mottagning av radio- och radarsignaler, där ett signal–brusförhållande analyseras. Fördelen med ROC-analys är att man kan analysera en metods prestanda när man har flera möjliga nivåer för gränsdragning mellan ”sjukt” och ”friskt”. I Figur B9.5 ges ett exempel på hur ROC kan användas. Rent matematiskt

är den optimala gränsdragningen mellan sjukt och friskt i den punkt som ligger närmast diagrammets övre vänstra hörn. Det bästa valet i en klinisk situation behöver dock inte vara där. Beroende på situationen, kan man välja att prioritera en stor andel sant positiva och acceptera att andelen falskt positiva blir relativt stor, eller välja en relativt låg andel sant positiva för att undvika en hög andel falskt positiva. Se tidigare diskussion. Om man jämför två testmetoder där den ena ligger helt ovanför den andra så är den förra helt klart bäst. Arean under ROC-kurvan är ett sammantaget mått på testets prestanda och är lika med sannolikheten att en slumpmässigt vald person med sjukdomen har ett högre värde än en slumpmässigt vald person utan sjukdomen. Arean 1 eller 100 procent anger att det är ett perfekt test och mindre än 0,5 eller 50 procent att det inte är bättre än slumpen.



Figur B9.5 ROC-kurvor.
”Receiver operating characteristic” (ROC) med sanna positiva värden (sensitivitet) på Y-axeln och falska positiva värden (1-specificitet) på X-axeln. Kurvorna visar sambandet mellan sant positiva värden och falskt positiva värden för dentinkaries (djup karies innanför emaljen) på approximal- (tandytor mot varandra) och ocklusalytor (tuggytor). Exemplet är från en in vitro-studie med extraherade tänder (bästa möjliga referensmetod = histologisk verifikation), där undersökare fick ange graden av säkerhet för dentinkaries på röntgenbilden på en skala från 0=”säkert ingen dentinkaries” till 5=”helt säkert dentinkaries” [7].

Reliabilitetsmått

Reliabiliteten (tillförlitligheten, precisionen) hos en diagnostisk metod är ett uttryck för hur väl t ex en kariesdiagnos överensstämmer mellan olika undersökare eller hur väl samma undersökare kan upprepa en specifik diagnos vid ett senare tillfälle. Överensstämmelsen inom eller mellan undersökare kan sammanfattas på olika sätt; som hur många procent av ett antal diagnoser man är överens om, som korrelationskoefficienter eller som kappavärde. Kappavärdet används ofta för att beskriva en diagnostisk metods reliabilitet. Kappavärdet är den observerade överensstämmelsen justerad för sannolikheten att överensstämmelsen beror på slumpen. I Exempel B9.8 redovisas ett exempel på hur kappavärdet beräknas.

Tumregler för att värdera kappavärdet har utvecklats, se Faktaruta B9.3. Det finns även andra mått för att bedöma överensstämmelsen som inte redovisas här.

Exempel B9.8 Beräkning av kappavärde: ett hypotetiskt exempel.

Röntgenbilder av 60 tandytor (approximalytor) tolkas av två undersökare med avseende på förekomst av djup karies (dentinkaries). Positiv = karies; negativ = ingen karies.

a) Resultatet av upprepad tolkning av röntgenbilder av 60 tandytor gav följande resultat:

	2:a tolkningen		
1:a tolkningen	**Positiv**	**Negativ**	**Totalt**
positiv	11	11	22
negativ	3	35	38
totalt	14	46	60

Vid första tolkningen hade 22 ytor karies och vid andra tolkningen som gjordes ”blint” (ovetande om vad observatören kommit fram till första gången) hade 14 ytor karies. Elva ytor var positiva båda gångerna och 35 ytor var negativa vid båda tillfällena. Den procentuella överensstämmelsen var alltså (11+35)/60=0,77 eller 77 procent. Detta mått på observatörsvariation är dock missledande, eftersom man ignorerar att de två tolkningarna kan ge samma resultat beroende på slumpen.

En del av överensstämmelsen hade även slumpmässigt kunna inträffa, dvs att båda tolkningarna gett positiva respektive negativa utfall.

Det är enkelt att beräkna det förväntade slumpmässiga resultatet om diagnoserna var oberoende av varandra. Om man tänker sig att observatören andra gången slumpmässigt väljer 14 röntgenbilder och benämner dem ”positiva”. Man kan då förvänta sig att 14x22/60=5,13 bilder kommer att bli ”positiva” vid båda tillfällena och att 46x38/60=29,13 kommer att bli ”negativa” båda gångerna. Det förväntade resultatet beroende på slumpen är illustrerat i b). Man kan alltså beräkna att det förväntade resultatet som beror på slumpen är (5,13+29,13)/60=0,57 eller 57 procent. När man tar detta i betraktande, blir den observerade procentuella överensstämmelsen på 77 procent mindre imponerande.

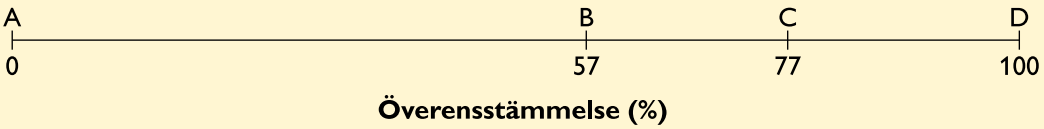
b) Det förväntade slumpmässiga resultatet om diagnoserna var oberoende av varandra.

	2:a tolkningen		
1:a tolkningen	**Positiv**	**Negativ**	**Totalt**
positiv	5,13	16,87	22,00
negativ	8,87	29,13	38,00
totalt	14,00	46,00	60,00

Exemplet fortsätter på nästa sida

Exempel B9.8 Fortsättning.

Det är detta problem som formaliserats i den så kallade kappastatistiken, som relaterar den observerade överensstämmelsen till den överensstämmelse som kan bero på slumpen. I det beskrivna exemplet är kappavärdet 47 procent, vilket betyder att skillnaden mellan den observerade överensstämmelsen och den slumpberoende överensstämmelsen (77–57 %) är endast 47 procent av skillnaden mellan perfekt överensstämmelse och slumpmässig överensstämmelse (100–57 %). Detta kan illustreras med en figur:



Kappa = (C–B)/(D–B) = (77–57)/(100–57) = 0,47 eller 47%.

A = total brist på överensstämmelse; B = den förväntade överensstämmelsen pga slumpen; C = den observerade överensstämmelsen; D = perfekt överensstämmelse

Faktaruta B9.3 Tumregler för att värdera kappavärdet.

Kappavärde	Grad av överensstämmelse
≤0,20	Dålig
0,21–0,40	Svag
0,41–0,60	Måttlig
0,61–0,80	Bra
0,81–1,00	Mycket bra

Övergripande diskussion om olika mätmetoders för- och nackdelar

Den här översikten visar att det finns många sätt att mäta diagnostiska testmetoders värde. Alla har sina för- och nackdelar. I vissa situationer är det viktigast att ha en hög sensitivitet, i andra att ha en hög specificitet. Sensitivitet och specificitet mäter testmetodens tillförlitlighet, men säger inte hur säkert man kan veta om patienterna som testats har sjukdomen eller ej. Det beror bl a på prevalensen av sjukdomen i den grupp som undersöks. I sådana sammanhang är positiva och negativa prediktionsvärden att föredra eftersom de tar hänsyn till såväl sensitivitet, specificitet som prevalens. ”Likelihood”-kvoter uttrycks ofta som odds och har fördelen att man sammanfattar sensitivitet och specificitet i ett mått som är oberoende av prevalensen. Å andra sidan påverkar alltid prevalensen i den grupp som studeras hur väl en specifik testmetod fungerar i praktiken. Att mäta oddset före och efter en testmetod kan vara ett bra sätt för att bedöma den marginella

nyttan av att ta ytterligare ett test. De grundläggande måtten är dock prevalens, sensitivitet och specificitet. Med hjälp av dessa tre grundmått kan alla andra mått beräknas.

Behandlingsstudier

Grundläggande statistiska mått när det gäller att bedöma utfallet av olika typer av behandlingsinsatser är:

- absoluta och relativa risker
- risk- och oddskvoter
- nödvändigt antal behandlingar för att förebygga ogynnsamma händelser (NNT)
- konfidensintervall och hantering av slumpen.

Effekten av en behandling kan uttryckas på många olika sätt och i såväl absoluta som relativa tal. Det finns för- och nackdelar med de flesta och man ska vara medveten om att vi tolkar data olika beroende på hur effekten uttrycks. Det finns alltså en risk att de som redovisar resultaten av en studie väljer mått som passar med det intryck de vill förmedla. Förutom grundläggande mått redovisas också hur man kan göra metaanalyser för att få en mer samlad bild av resultaten.

De statistiska mått som används och redovisas nedan är desamma för samtliga studietyper. Behovet av att kontrollera för bakomliggande faktorer är dock mycket större för kohort- och fall–kontrollstudier. Ett sätt att statistiskt hantera sådana metodproblem är att stratifiera analyserna i olika undergrupper eller göra regressionsanalyser.

Utgångspunkten för beräkning av olika statistiska mått är en fyrfältstabell där vi på raderna har de två behandlingsinsatser eller metoder som jämförs, dvs en försöks- och en kontrollgrupp (Exempel B9.9). I kolumnerna anges antalet personer och antal utfall/händelser i respektive grupp. Utfallet kan vara dödsfall, sjukdomsincidens etc. Med denna tabell kan olika riskmått beräknas.

Absoluta och relativa risker

Riskmått kan uttryckas som absolut risk och relativ risk för en händelse eller ett utfall som är förknippat med åtgärden/interventionen i fråga. Vid en jämförelse mellan försöks- och kontrollgrupp talar man om absolut och relativ riskreduktion. Utifrån fyrfältstabellen kan vi definiera begreppen och illustrera beräkningarna med ett fiktivt exempel, se Exempel B9.9.

Fördelen med ett absolut riskmått är att det anger den absoluta risken för att en viss händelse ska inträffa utan att jämföra risken med någon annan. Eftersom det lyckligtvis

är så att negativa utfall är relativt sällsynta blir riskreduktionen ofta väldigt låg. Den relativa riskreduktionen visar effekten i relation till ett alternativ och blir därför högre. Av förklarliga skäl föredrar industrin som har kommersiella intressen oftast att redovisa resultaten i relativa termer eftersom det förstärker de positiva effekterna. Effekten upplevs som större.

Exempel B9.9 Absoluta och relativa risker.

Studiegrupper	Antal personer	Antal utfall/händelser
Försöksgrupp, t ex läkemedel A	a) 144	c) 19
Kontrollgrupp, t ex placebo, annat läkemedel	b) 148	d) 24

Absolut risk i försöksgruppen = c/a = 19/144 = 0,132 = 13,2 %.

Absolut risk i kontrollgruppen = d/b = 24/148 = 0,162 = 16,2 %.

Absolut riskreduktion = absolut risk i kontrollgruppen – absolut risk i försöksgruppen = 16,2 – 13,2 = 3,0 procentenheter.

Relativ riskreduktion = absolut riskreduktion/absolut risk i kontrollgruppen = 3,0/16,2 = 18,5 %.

Risk- och oddskvoter

De två grundläggande måtten när man jämför två åtgärder eller behandlingsalternativ är relativ risk (RR) och oddskvot (”odds ratio”, OR). Definition och räkneexempel presenteras i Faktaruta B9.4 och Exempel B9.10.

Faktaruta B9.4 Formler för relativ risk och oddskvot.

Relativ risk (RR) = absolut risk i försöksgruppen/absolut risk i kontrollgruppen.

Odds = sannolikhet att ha utfallet (t ex avlida)/sannolikhet att inte ha utfallet (t ex leva).

Oddskvot (OR) = odds i försöksgruppen/odds i kontrollgruppen.

För de flesta är det svårare att tolka en oddskvot än en riskkvot. Skillnaden mellan odds och risk blir mindre ju mer sällsynt händelsen är. Om oddset är 1 mot 10 eller 0,1 blir risken 0,91 (1 av 11). När händelserna är vanliga blir skillnaden mellan odds och risk stor. En risk på 0,5 är detsamma som ett odds som är 1 och en risk på 0,95 är detsamma som ett odds på 19.

Exempel B9.10 Risk- och oddskvoter.

Studiegrupper	Antal personer	Antal utfall/ händelser	Risk	Odds
Försöksgrupp, t ex läkemedel A	a) 144	c) 19	19/144=0,132	19/(144–19)=0,152
Kontrollgrupp, t ex placebo, annat läkemedel	b) 148	d) 24	24/148=0,162	24/(148–24)=0,194

Relativ risk (RR) = 0,132/0,162 = 0,81.

Oddskvot (OR) = 0,152/0,194 = 0,78.

Genom multiplikation med 100 omformas riskkvoter (men inte oddskvoter) till procenttal. Ofta uttrycker man effekten i termer av riskminskning. Om riskkvoten är 0,25 blir riskminskningen 75 procent, dvs 100 x (1–0,25). Den kliniska betydelsen av en riskminskning kan ses vid jämförelse med den absoluta risken i kontrollgruppen. Om händelsen drabbar 80 procent i kontrollgruppen och 60 procent i experimentgruppen kan det ha en helt annan innebörd än om händelsen drabbar 4 procent i kontrollgruppen och 3 procent i experimentgruppen.

Det blir problem om en oddskvot tolkas som en riskkvot. Vid åtgärder som ökar sannolikheten för händelser kommer oddskvoten att bli större än riskkvoten, särskilt när händelsen är vanlig, varför feltolkning leder till överskattning av behandlingseffekten. För åtgärder som minskar sannolikheten för händelser kommer oddskvoten att bli mindre än riskkvoten, så även i denna situation är det lätt att misstolka resultatet, vilket är ganska vanligt.

Oddskvoter kan omvandlas till riskkvoten och riskkvoter till oddskvoter. Detta sker genom jämförelse med risken i kontrollgruppen (RK) (eller genom något annat mått på grundrisk). Det görs på följande sätt:

$$RR = \frac{OR}{1–RK(1–OR)} \quad \text{medför} \quad OR = \frac{RR(1–RK)}{1–(RK \times RR)}$$

Dessa omvandlingar kan också behövas när man i olika studier ömsevis använt det ena eller andra måttet.

Nödvändiga antal behandlingar för att förebygga ogynnsamma händelser (NNT)

För att få en intuitiv uppfattning om möjligheterna att hjälpa patienter brukar man beräkna det antal patienter som måste behandlas för att förebygga en ogynnsam händelse, t ex en hjärtinfarkt, död eller återfall i cancer. Detta mått kallas för nödvändiga antal behandlingar (”number needed to treat”, NNT) och beräknas som det inverterade värdet av den absoluta riskreduktionen. Om den absoluta riskreduktionen är 3 procent så blir NNT=1/0,03~34, dvs i genomsnitt krävs det att man behandlar 34 patienter för att undvika en ogynnsam händelse som död.

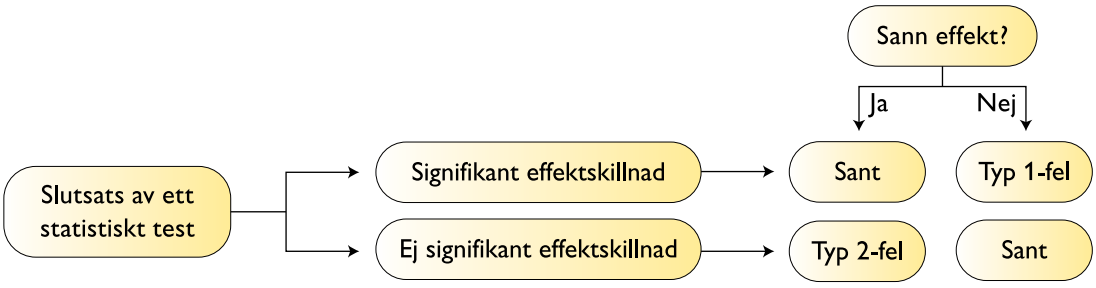
Exempel B9.11 NNT för att undvika en stroke.

I en metaanalys av sju studier där man jämförde betablockadsbehandling av mild till måttlig hypertoni med placebo var den absoluta riskreduktionen i stroke 0,22 procent [8]. Det innebär att 455 patienter (1/0,0022) behöver behandlas under 3–5 år för att undvika en stroke.

När det gäller risker går det att beräkna ett motsvarande mått, dvs nödvändiga antal behandlingar för att en skada ska uppkomma (”number needed to harm”, NNH).

Konfidensintervall och hantering av slumpen

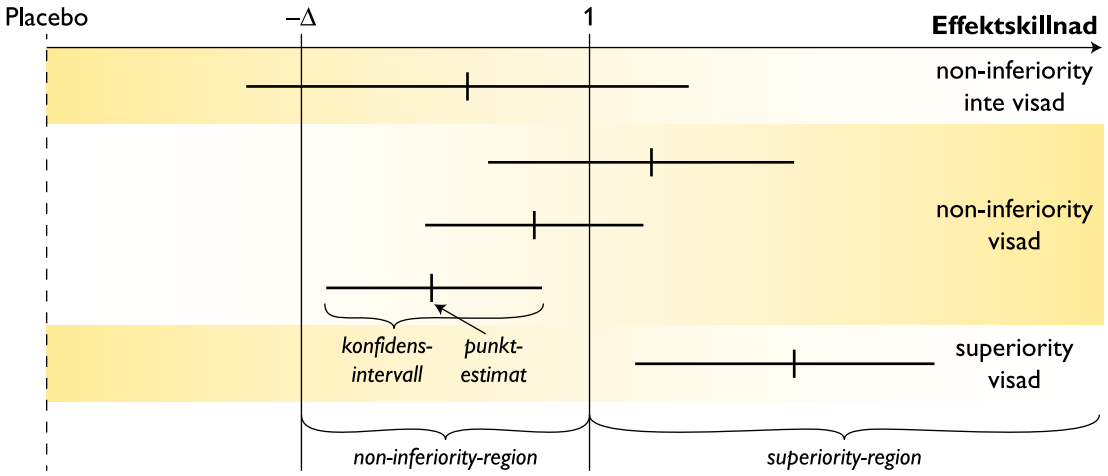
Slumpen kan göra att de resultat vi finner inte är sanna. Det finns två sätt att göra fel med en statistisk metod. Antingen kan man dra en falskt positiv slutsats, dvs att en ny behandlingsmetod är effektiv när den inte är det. Denna typ av felaktig slutsats kallas Typ 1-fel eller α-fel och definieras som sannolikheten att man ser en skillnad i behandlingseffekt när det inte finns någon (Figur B9.6). Den andra typen av fel är när man drar slutsatsen att den nya behandlingsmetoden inte är effektivare fast den i verkligheten är det. Man drar då en falskt negativ slutsats vilket kallas Typ 2-fel eller β-fel.



Figur B9.6 Samband mellan resultaten av ett statistiskt test och den sanna effekten mellan två behandlingsinsatser.

Konfidensintervallen visar på den statistiska osäkerheten i det urval som studeras. De bygger oftast på att man accepterar att Typ 1-fel (α) är mindre än 0,05. Små urval ger stora konfidensintervall. För beräkning av konfidensintervall hänvisas till statistiska metodböcker [1–3]. Vanligtvis väljs 95-procentiga eller 99-procentiga konfidensintervall. Med ett 95-procentigt konfidensintervall kommer det skattade värdet i 95 av 100 fall att hamna inom konfidensintervallet. Det är viktigt att inse att vi talar om sannolikhetsbedömningar och inte om en sanning. Även om resultaten inte är statistiskt signifikanta så kan det finnas ett verkligt samband.

Resultaten kan presenteras som i Figur B9.7. På engelska kallas det för en ”forest plot” där mittlinjen 1 indikerar att de två interventionerna (åtgärderna) är likvärdiga. Om hela konfidensintervallet för oddskvoten eller den relativa risken ligger under 1 så är den studerade interventionen statistiskt säkerställt bättre. Ligger konfidensintervallet för OR/RR helt över 1 är interventionen statistiskt säkerställt sämre. Korsar konfidensintervallet linjen 1 har vi ingen statistiskt säkerställd skillnad. Punkten i mitten anger det uppskattade värdet (punktestimatet).



Figur B9.7 Punktestimat och konfidensintervall för olika utfall av effektskillnader mellan kontrollbehandling (=1) och experimentbehandling i en RCT. Konfidensintervallets nedre gräns avgör om resultatet är förenligt med ”superiority” respektive ”non-inferiority”.

”Superiority” vs ”non-inferiority”-studier

Studier som visar klart bättre effekt för en behandling (överlägsenhet, ”superiority”) har ett inbyggt kvalitetsmått om randomiseringen är korrekt och man kan utesluta bias (brister i blindning). Brister i studiens kvalitet i övrigt kan möjligen göra att eventuella effektskillnader mellan behandlingsarmarna underskattas. En förutsättning är dock att behandlingen i kontrollgruppen är optimal, exempelvis att dosval i jämförelsegrupperna i läkemedelsprövningar varit rättvist.

Ibland är det dock lämpligt att designa en studie för att visa att det inte föreligger någon (kliniskt relevant) skillnad (”non-inferiority”). Till exempel är det i vissa situationer oetiskt att använda placebokontroll samtidigt som man ibland inte kan förvänta sig att ett läkemedel visar bättre effekt än standardbehandling. Om ett läkemedel/metod har en klart bättre säkerhetsprofil än standardbehandling men sannolikt inte är effektivare är en ”non-inferiority”-design nödvändig. När målet att visa en behandlings överlägsenhet inte uppnås, kan resultatet bedömas som ”non-inferior” om prövningens kvalitet tillåter detta [9].

Vid granskning av ”non-inferiority”-studier är det viktigt att bedöma vad som är en rimlig kliniskt relevant effektskillnad och underlaget för prövarnas uppskattning av detta förspecificerade delta (Δ). Valet av Δ måste uppfylla två krav:

1. Det krävs att Δ är specificerat så att man kan vara rimligt övertygad om att Δ klart skiljer sig från placebo beaktande osäkerhet rörande förändringar i övriga faktorer som över tid förbättrat prognosen av tillståndet och därmed påverkar effekten. Som ett riktmärke är det rimligt att kräva att minst 50 procent av referensbehandlingens effekt kvarstår (Figur B9.7).

2. Valet av Δ skall återspegla en rimlig klinisk uppfattning av vad som är relevant effektskillnad.

Vid beräkning av Δ enligt punkt 1 ovan är det viktigt att man utgår från en eller helst flera RCT som jämfört kontrollbehandlingen med placebo (alternativt annan kontroll vid hårda utfallsvariabler) under samma betingelser (patientgrupp, dos, stadium, utfallsvariabel, responskriterier, etc) som den aktuella studien. Dessa bör heller inte vara alltför gamla eftersom nya undersökningsmetoder, samtidig behandling och kontrollbehandlingens relevans påverkas av utvecklingen.

Det prespecificerade Δ är ett planeringsinstrument för att dimensionera studien, den finala ”non-inferiority”-värderingen är en nytta/risk-värdering av det observerade utfallet vid vilken man inte är bunden vid det *a piori* uppsatta Δ. Vad som kan anses utgöra en rimlig klinisk relevans (se nedan) är således inte bara en effektfråga utan är också avhängig respektive behandlings säkerhetsprofil. En klar skillnad i biverkningsmönster till fördel för experimentarmen kan göra det rimligt att acceptera ett större Δ.

Vidare är bedömning av datakvalitet, studiens upplägg och genomförande fundamental vid granskningen av en ”non-inferiority”-studie. Brister på inom dessa områden (patientpopulation/diagnostik, mätmetoder, mätfrekvens, val av dos, omhändertagande- och andra studieeffekter) tenderar alla att försämra möjligheten att upptäcka verkliga skillnader mellan behandlingar och därmed ”simulera” likhet. Även stora bortfall gynnar likhet

varför såväl per protokollanalys som ITT-analys också krävs för att ”non-inferiority” ska accepteras [10].

Klinisk relevans

Klinisk relevans är ett lika viktigt som svårdefinierat begrepp. En kliniskt relevant effekt är den effekt som är meningsfull att upptäcka alternativt utesluta, beroende på studiedesign. Begreppet är grundläggande i planeringen av jämförande studier (”power” i ”superiority”-studier och deltaberäkning i ”non-inferiority”-studier). Även storleken på fas II-studier bestäms utifrån antagande om vad som är en kliniskt relevant effekt vilken är värd att upptäcka. Dessutom är det krav vid godkännande av ett nytt läkemedel att man kunnat demonstrera en kliniskt relevant effekt. Vid värderingen av de observerade resultaten ska man inte fästa sig vid vad som specificerats vid beräkning av den nödvändiga studiestorleken. Denna beräkning går ut på att få en hög sannolikhet (”power”) för ett statistiskt signifikant resultat givet att den specificerade skillnaden är sann. Den garanterar inte hög sannolikhet för att observera en skillnad som är minst lika stor som den specificerade. Sannolikheten för detta är bara 50 procent givet att antagandet är sant. Med hög ”power” kommer även mindre observerade skillnader att bli statistiskt säkerställda och det är dessa som ska bedömas vägt mot observerade och ej observerade potentiella biverkningar för att avgöra om resultatet är kliniskt relevant.

Vad som är en kliniskt relevant effekt beror i första hand på vems perspektiv man utgår ifrån; patientens, anhörigas, behandlande läkare, myndighet eller ”third party payers”. Åtskilliga undersökningar har kunnat slå fast att vad som uppfattas som en relevant effekt i relation till risker och kostnader kraftigt påverkas av om man är föremål för åtgärden, ordinatör eller bara ”bystander”.

Exempel B9.12 Behandlingars värde och klinisk relevans.

I en studie av attityder till kemoterapi vid behandling av cancer tillfrågades patienter, friska kontroller, specialistläkare, allmänläkare och sköterskor om vilket behandlingsalternativ de skulle välja i olika situationer. Resultatet visade att patienterna var mycket mer riskbenägna och villiga att acceptera biverkningar än övriga grupperna, trots låga odds för tillfrisknande [11]. Den här typen av information är av stor vikt vid diskussioner om behandlingars värde och klinisk relevans.

Naturligtvis varierar uppfattningen kraftigt även mellan individer inom en grupp och sannolikt också från ett tillfälle till ett annat och mellan olika perioder i livet hos samma individ. Exempelvis har sjuka individers förmåga att hantera sin sjukdom stor betydelse för förändringar över tid när det gäller uppfattningen om klinisk relevans. Det är dock

rimligt att anta att för en enskild individ uppfattas en behandling med stor effekt som mer relevant än en behandling med liten effekt, allt annat lika.

En effekt kan i praktiken vara mer eller mindre relevant (kontinuerlig variabel). Oftast hanterar man dock klinisk relevans som en dikotom variabel, dvs man önskar fastställa vad som är en relevant respektive icke relevant klinisk effekt. Detta är en praktisk förenkling av verkligheten på samma sätt som frågan om vem som kan anses vara responder eller ej.

Medan statistisk signifikans är lätt att definiera, närmast av axiomatisk natur och ett begrepp som det skrivits tusentals vetenskapliga publikationer om, är litteratur omkring begreppet klinisk relevans sparsam. Klinisk relevans går inte att slå fast utifrån kvantitativa metoder även om den till syvende och sist måste anges med ett kvantitativt mått. Det är snarare kvalitativa data som kan bibringa en uppfattning om vad som är en kliniskt relevant effekt. Dessa blir dock alltid föremål för en subjektiv bedömning vilken ånyo är avhängig bedömarens perspektiv, dvs vilken grupp man tillhör.

Patientens upplevelser kan i allmänhet fångas på tre nivåer; i form av ”patient satisfaction”, som hälsorelaterad livskvalitet (HRQoL) eller i form av kvalitativ forskning, t ex intervjuer.

- Det är viktigt att vara tydlig med skillnad mellan registrering av ”patient satisfaction” vilket är ett behandlingsutfall som kan vara kopplat till HRQoL (men inte nödvändigtvis behöver vara det) och HRQoL.

- HRQoL är index som mer övergripande påverkas av individens hela livssituation, det som betecknas som individens livsvärld, och som i sig inrymmer psyklogiska, sociala och medicinska aspekter som av många anledningar fluktuerar över tid för alla individer.

- Forskningsintervjuer är ett mänskligt samtal med syfte att få en beskrivning av den intervjuade för att beskriva och/eller tolka upplevelser/innebörd eller mening i förhållande till den intervjuades livsvärld. Intervjuerna genomförs som ett samtal omkring upplevelserna där den intervjuade fritt berättar och där intervjuarens uppgift är att ställa stödfrågor om så behövs.

För att fånga patientens upplevelse av den kliniska relevansen behövs sannolikt alla tre nivåerna. Troligen kommer dock den mest relevanta informationen fram i forskningsintervjun.

Vad som uppfattas som en kliniskt relevant effekt är också avhängigt en behandlings säkerhetsprofil. Effekten av en behandling med små risker kan uppfattas som kliniskt relevant medan samma effekt hos en behandling med uttalade risker kan uppfattas som varande icke relevant. Med andra ord borde man snarare tala om kliniskt relevant ”nytta/risk-balans” snarare än kliniskt relevant effekt.

När det rör sig om symtomatisk behandling av benigna sjukdomar där otillräcklig behandling inte försvårar prognosen på sikt är det rimligt att lämna bedömningen om storleken av effekt i relation till frekvens av biverkningar av toleranskaraktär till patienten. Om det däremot finns allvarligare säkerhetsproblem, empiriskt identifierade eller potentiella utifrån prekliniska data, verkningsmekanism, klasseffekter etc, är det i många fall rimligt att nytta/risk-värderingen görs av någon annan än patienten.

Vid allvarliga sjukdomar där det troliga utfallet är död är förhållandet det motsatta. Frekvent intolerans som påverkar livskvalitet blir viktig, särskilt om effektfördelen rör sig om små skillnader i tid till död i sjukdomen. Däremot är allvarliga infrekventa biverkningar mindre viktiga i denna situation, i varje fall om risken för dem är klart mindre än risken för död i sjukdomen.

Klinisk relevans ska inte blandas ihop med kostnadseffektivitet även om det är ett grundvillkor för att en metod ska kunna anses kostnadseffektiv att dess effekt bedömts vara kliniskt relevant, medan det omvända inte gäller.

Den politiska dimensionen handlar om vem (och var) som ska bedöma vad som är en kliniskt relevant effekt. Fördelen med att låta den välinformerade patienten i varje situation själv eller i samråd med sin behandlare bedöma vad som är kliniskt relevant är att den ovan beskrivna inter- och intraindividuella variabiliteten då har hanterats. Nackdelen är att rättighetsaspekter (jämlik vård) inte tillgodosetts alls. Omvänt är fördelen med en myndighetsbedömning att rättviseaspekten och frågan om optimalt resursutnyttjande möjligen hanterats bättre. Dessutom har inte alla patienter förmåga att ta till sig information så att de kan anses vara välinformerade. Därmed måste frågan om klinisk relevans i många fall ändå hanteras av behandlare, högre administrativ nivå eller av myndighet. Det är i denna situation viktigt att beslutsfattaren är väl införstådd med målgruppens uppfattning om klinisk relevans.

Referenser

1. Altman DG. Practical statistics for medical research. Chapman and Hall/Crc Pree Llc, 1991.

2. Bland M. An introduction to medical statistics. Oxford University Press, 3rd ed, 2000.

3. Bring J, Taube A. Introduktion till medicinsk statistik. Studentlitteratur, 2006.

4. Taube A, Malmquist J. Räkna med vad du tror. Bayes' sats i diagnostiken. Läkartidningen 2001;98:2910-3.

5. Taube A, Malmquist J. Räkna med vad du tror. Bayes – inte P-värdet – mäter tilltron. Läkartidningen 2001;98:3208-11.

6. Bytzer P, Hansen JM, Schaffalitzky de Muckadell OB. Empirical H2-blocker therapy or prompt endoscopy in management of dyspepsia. Lancet 1984;343:811-6.

7. Espelid I, Tveit AB. A comparison of radiographic occlusal and approximal caries diagnoses made by 240 dentists. Acta Odontol Scand 2001;59:285-9.

8. Lindholm LH, Carlberg B, Samuelsson O. Should β blockers remain first choice in the treatment of primary hypertension? A meta-anlysis. Lancet 2005;366:1545-53.

9. Guideline on the choice of the non-inferiority margin. http://www.ema.europa.eu/pdfs/human/ewp/215899en.pdf

10. Points to consider on switching between superiority and non-inferiority. http://www.ema.europa.eu/pdfs/human/ewp/048299en.pdf

11. Slevin ML, Stubbs L, Plant HJ, Wilson P, Gregory WM, Armes PJ, Downer SM. Attitudes to chemotherapy: comparing views of patients with cancer with those of doctors, nurses, and general public. BMJ 1990;300:1458-60.

SBU ger kunskap för en bättre vård

Medicin och hälsa är viktiga områden som berör oss alla. Som medborgare och patienter vill vi ha tillgång till den bästa möjliga vården – men hur vet vi att de metoder vi använder är de säkraste och mest effektiva?

Inom hälso- och sjukvården är personalen enligt lag skyldig att arbeta enligt vetenskap och beprövad erfarenhet. Samtidigt publiceras en så stor mängd vetenskapliga artiklar att det blir omöjligt för den enskilde vårdgivaren att hinna följa med i den ständigt växande strömmen av nya forskningsrön. Forskningsresultaten behöver sorteras, granskas kritiskt och sammanställas så att de blir överskådliga.

SBU, Statens beredning för medicinsk utvärdering, har i uppdrag att utvärdera metoder som används i vården, både etablerade och nya. Utifrån aktuell och välgjord forskning tar vi reda på vilken medicinsk effekt olika metoder har, om det finns några risker med dem, och om åtgärderna ger mesta möjliga nytta för pengarna.

SBU:s oberoende utvärderingar ska användas som stöd av alla som på olika nivåer i samhället bestämmer hur hälso- och sjukvården ska se ut. Vi kan peka på möjligheter till ytterligare förbättring, så att sjukvården kan använda sina resurser på bästa sätt och Sveriges befolkning kan få en bättre hälsa. SBU ger tillförlitliga svar på sådana frågor. Vi är en oberoende statlig myndighet som har regeringens uppdrag att utvärdera vårdens metoder ur ett samlat medicinskt, ekonomiskt, etiskt och socialt perspektiv.

Våra utvärderingar är så kallade systematiska litteraturöversikter, som bygger på publicerad forskning och följer en väl genomarbetad och noggrann metod.

SBU har gjort systematiska översikter sedan 1987, och det gör oss till en av de äldsta HTA-organisationerna (Health Technology Assessment) i världen.

Läs SBU:s utvärderingar på www.sbu.se.