SBU:S METODBOK
PUBLICERAD: 15 OKTOBER 2020
NEDLADDAD: 18 NOVEMBER 2020

SBU:s metodbok

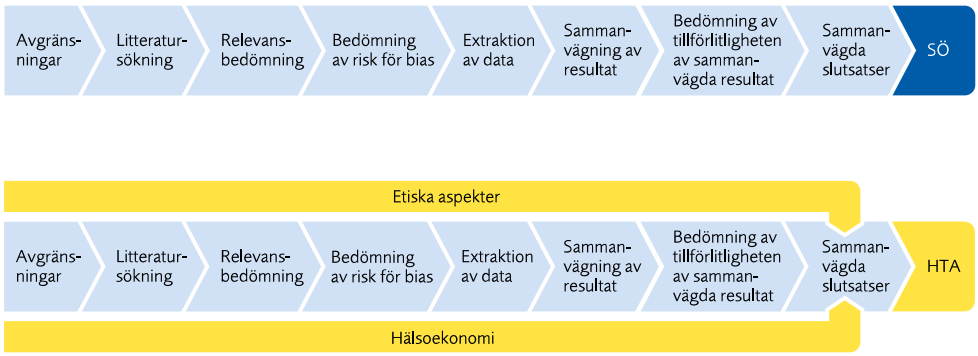
STATENS BEREDNING FÖR MEDICINSK OCH SOCIAL UTVÄRDERING

Innehåll

1. Introduktion

Det är viktigt att de interventioner och metoder som används i hälso- och sjukvården och socialtjänsten har vetenskapligt stöd. SBU:s uppdrag är att utvärdera metoder som används eller kan användas inom olika sektorer av hälso- och sjukvården och socialtjänsten. Grunden är en systematisk översikt, som kan vara utförd av SBU eller av någon annan organisation eller forskargrupp. En fullständig utvärdering, en så kallad Health Technology Assessment (HTA), omfattar även hälsoekonomi samt bedömningar av etiska och sociala aspekter som har betydelse för användning av metoden, se Figur 1.

Figur 1 Blå-tonade pilar visar vilka delar som ingår både i en systematisk översikt (SÖ) och i en HTA-rapport (eng. Health technology assessment, HTA). Kompletterar man den systematiska översikten med utvärdering av hälsoekonomiska och etiska aspekter uppfyller den kraven för en HTA-rapport.



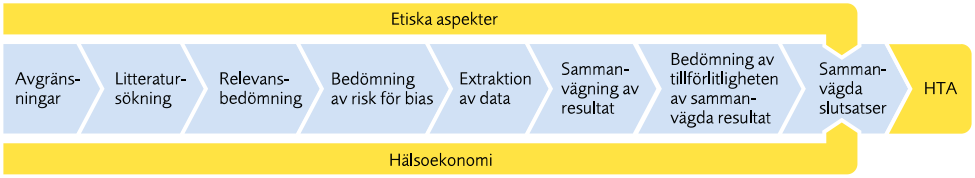
Denna metodbok är tänkt som ett praktiskt stöd för dem som ska genomföra en utvärdering och ersätter därmed inte de läroböcker som idag finns om systematiska översikter.

Metodbokens första del rör de olika stegen i en systematisk översikt med separata kapitel för formulering av forskningsfrågorna, identifiering av litteratur, granskning av enskilda studier, syntes av resultaten från enskilda studier och bedömning av hur tillförlitliga resultaten av syntesen är (tidigare kallat ”evidensstyrka”). I en andra del beskrivs principerna för att använda andras systematiska översikter, helt eller delvis. Metodboken avslutas med hur SBU arbetar med de delar som behövs för att producera en komplett HTA-rapport, det vill säga hälsoekonomi samt etiska aspekter.

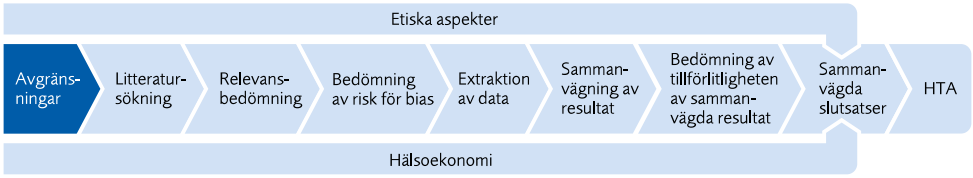
2. Processen för att ta fram en systematisk översikt

En systematisk översikt (eng. *systematic review*) ska uppfylla höga krav på tillförlitlighet och ska följa principer som minimerar risker för att slump, systematiska fel eller subjektiva värderingar påverkar slutsatserna. Den ska också rapporteras på ett sådant sätt att läsaren har en möjlighet att granska hur översikten har tagits fram. SBU följer de internationella riktlinjer som finns för hur systematiska översikter ska genomföras och rapporteras. Preferred Reported Items for Systematic Reviews and Meta-analyses (PRISMA) [1] är främst avsedd för studier med kvantitativ metodik. Det finns flera versioner av PRISMA, till exempel för interventionsstudier och för studier om diagnostisk tillförlitlighet [2]. För systematiska översikter som bygger på studier med kvalitativ metodik finns riktlinjerna ”The Enhancing transparency in reporting the synthesis of qualitative research” (ENTREQ) [3]. De aktuella versionerna av båda riktlinjerna finns på webbplatsen för det internationella nätverket EQUATOR network.

Figur 2 Steg i att producera en HTA-rapport. Blåtonade rutor visar vilka delar som ingår i en systematisk översikt.



3. Avgränsningar för den systematiska översikten



Arbetet med en systematisk översikt inleds med att skriva en projektplan eller protokoll för den systematiska översikten. Den ska beskriva syftet och frågor som ska besvaras med den systematiska översikten samt arbetssättet. Det finns en mall för hur projektplanen för ett SBU-projekt ska se ut. Projektplanen bör publiceras i den internationella databasen PROSPERO som innehåller protokoll för systematiska översikter. Detta är nödvändigt om översikten även ska publiceras i vetenskaplig tidskrift.

Den forskningsfråga som översikten ska besvara måste vara fokuserad och strukturerad. En ostrukturerad fråga leder till problem genom hela processen, till exempel svårigheter att skapa bästa möjliga sökstrategier och att bedöma vilka studier som är relevanta. Första steget i processen är därför att specificera frågan med stöd av ett strukturerat format (Faktaruta 3.1).

Faktaruta 3.1 Strukturerat frågeformat för olika typer av frågor.

Fråga	Strukturerat format	Betydelse
Effekt av intervention	PICO	Population Intervention Kontroll (Control) Utfall (Outcome)
Samband	PECO	Population Exponering Kontroll (Control) Utfall (Outcome)
Diagnostisk tillförlitlighet	PIRO	Population Indextest Referenstest Utfall (Outcome)
Erfarenheter och värderingar	SPICE	Setting Population Intervention Jämförelse (Comparison) Utfall (Evaluation)

Dessa standardiserade format kommer i praktiken att fungera som inklusions- och exklusionskriterier. Det är viktigt att vara medveten om konsekvenserna av att avgränsa frågan med till exempel PICO. De studier som uppfyller PICO kommer att inkluderas medan övriga studier kommer att exkluderas.

Detta kapitel beskriver kortfattat de format som SBU använder för att ställa strukturerade frågor om effekter av interventioner, om effekter av exponering, om värdet av (diagnostiska) tester samt om erfarenheter och upplevelser. Dessutom beskrivs några andra kriterier som ofta behövs för att avgränsa frågan ytterligare.

3.1 Format för effekter av behandlingar och interventioner

Behandlingar och insatser kommer fortsättningsvis att kallas interventioner, för att anpassa till den internationellt vedertagna terminologin.

Som nämndes ovan används begreppet PICO (på svenska: Population, Intervention, Kontroll, Utfall) för att strukturera frågor om effekter av behandlingar eller interventioner. PICO innebär att fyra delar av frågan ska specificeras: för vilka patienter eller personer är det relevant att undersöka effekten av interventionen? Vilken är interventionen? Vad kan fungera som kontrollåtgärd? Vilka positiva och negativa utfall av interventionen är vi intresserade av och hur ska de mätas?

För att få ett så bra PICO som möjligt bör såväl kliniker, som patienter och brukare involveras.

3.1.1 Population

Generellt bör populationen definieras så specifikt som möjligt. Gäller frågan både kvinnor och män? Ska effekten utvärderas för specifika åldersgrupper? Hur stringenta krav ska ställas på diagnos? Accepteras bara studier som använt diagnostiska kriterier och i så fall vilka kriterier? Räcker det med självrapporterade problem? Finns det någon samsjuklighet att ta hänsyn till? Ska frågan enbart gälla personer som brukar eller missbrukar någon drog?

Ett problem är hur studier som inte haft samma avgränsningar som den systematiska översikten ska hanteras. Ett exempel är effekter av interventioner till äldre, som är ett diffust begrepp. Här måste först den (nedre) åldersgränsen specificeras, till exempel över 65 år eller över 80 år. Därefter behövs kriterier för hur stor andel av deltagarna som uppfyller ålderskravet, något som kräver att studien redovisar åldersfördelningen. Observera att det ibland kan finnas en subgruppsanalys på åldersgrupper som bara redovisas i ett appendix på tidskriftens webbsida. Läs mer nedan.

Mer om population

Man bör redan i projektplanen specificera om den primära populationen är en subgrupp. Om det finns en tydlig effekt i den primära populationen är det rimligt att undersöka om effekten gäller hela populationen, eller om den drivs av effekten i en begränsad subgrupp.

Det är inte självklart vilka krav man ska ställa på statistisk signifikansnivå i sådana subgruppsanalyser eftersom studier ofta inte är dimensionerade för att utesluta skillnader mellan olika subgrupper. Dessutom är konsekvensen av ett falskt positivt utfall i en subgrupp i denna situation mindre allvarlig än vid testning av grundhypotesen. Frågan om signifikansnivå i subgruppsanalyser kan därför med fördel tas upp i rapportens diskussionsdel.

Om däremot utfallet i den primära populationen inte visar någon statistiskt säkerställd effekt, bör man vara mycket återhållsam med subgruppsanalyser såvida dessa inte varit förspecificerade och den statistiska signifikansnivån (α) justerats för multipla signifikanstester. Det är egentligen ointressant hur primärstudierna har hanterat subgruppsanalyser så länge som man i den egna rapporten håller sig till sitt definierade PICO och att populationen är definierad utan kännedom om eventuella utfall i subgrupper i primärstudierna.

3.1.2 Intervention

Interventioner är ett brett begrepp och omfattar såväl medicinska och odontologiska behandlingar som interventioner som ges i socialtjänsten eller inom funktionshinderområdet. Effekter av att använda diagnostiska metoder räknas också hit, till exempel om tillägg av ett test ökar överlevnaden. Interventionen kan utgöras av en enstaka åtgärd men också omfatta flera komponenter, så kallade komplexa interventioner.

Även här kan avgränsningarna göras mer eller mindre snäva. Dos, intensitet, hur länge intervetnionen ska pågå och vem som ska administrera interventionen kan vara faktorer som behöver specificeras.

3.1.3 Kontroll

Effekten av interventionen ska jämföras med effekten av en annan intervention, kontrollen. Det är viktigt att kontrollen är rimlig och ges under rimliga förhållanden. Det finns några huvudtyper av kontroller och PICO kan omfatta både en eller flera av dem.

En vanligt förekommande kontroll är att deltagarna inte får någon intervention utöver sedvanliga rutiner: sedvanlig behandling, sedvanlig handläggning, sedvanligt skolschema med mera. Nackdelen med sedvanliga rutiner är att de oftast inte är definierade och därmed kan variera mellan exempelvis olika kliniker, länder och hälso- och sjukvårdssystem. De sedvanliga rutinerna kan också förändras över tid om till exempel nya rutiner eller standardbehandlingar införs. En risk är att den intervention som studeras införs även för kontrollgruppen vilket medför att en eventuell effekt av interventionen minskar på grund av kontaminering. Ett alternativ till sedvanlig behandling kan vara att använda sig av väntelista där de som tilldelas kontrollinterventionen får den experimentella interventionen efter en viss tid. En nackdel är att det då inte går att göra några jämförande långtidsuppföljningar.

En annan variant är att deltagarna utgör sina egna kontroller. Några exempel är fallskydd för att förebygga höftfrakturer där den ena höften är skyddad men inte

den andra, eller behandlingar i munnen där ena halvan av munnen får den experimentella metoden och den andra inte.

En annan vanlig kontroll är att deltagarna får en intervention som anses vara verkningslös, till exempel placebo för läkemedel och ostrukturerade samtal för psykologiska behandlingar.

För interventioner där det redan finns tillgängliga metoder med vetenskapligt stöd kan det vara värdefullt att jämföra den nya interventionen med en etablerad. Det gäller då att vara uppmärksam på att kontrollinterventionen ges i rimlig dos. Det finns exempel på att den experimentella interventionen ges under optimala betingelser medan jämförelsen ges i lägsta dos.

3.1.4 Utfall

Här specificerar vi vilka utfall, det vill säga vilka effekter av interventionen, som vi är intresserade av. Utfall kan vara av olika betydelse för dem som interventionen riktas till. De kan vara kritiska (t.ex. minskad dödlighet) eller viktiga (t.ex. minskad grad av problem eller ökad livskvalitet). Vissa utfall, som till exempel laboratorievärden, är endast viktiga om de har en direkt koppling till hälsoutfall. I övrigt kan de ses som mindre viktiga.

Man bör inte ha för många utfall. Förslagsvis definieras *ett* primärt utfall och därefter ett, eller några få, sekundära mått. Det primära utfallet i en systematisk översikt bör vara kritiskt eller mycket viktigt för patienten eller brukaren. Negativa konsekvenser av en intervention, till exempel komplikationer eller ökning av problembeteenden ska alltid ingå, liksom kostnadseffektivitet om projektet är en HTA-rapport.

Specificeringen av utfall omfattar även *hur* de ska mätas, med *vilket mått*, samt *när*. Mätmetoderna ska vara validerade och reliabla. Utfallet kan mätas vid en eller flera tidpunkter, varav en är primär. Ett exempel är mätning av effekter av preventiva interventioner där den önskvärda effekten kan ligga decennier framåt i tiden. Sådana långa uppföljningstider är sällsynta. Många studier nöjer sig med att mäta effekten direkt efter avslutad intervention eller några månader senare. Risken är att uppmätt effekt då snarare kan ses som effekter av en behandling än som effekter av prevention. Internationella riktlinjer rekommenderar att preventiva interventioner följs upp minst sex månader efter avslutad intervention [4].

Kompositmått är också vanligt förekommande i klinisk forskning. Det innebär att man räknar samman flera olika effektmått för ett specifikt utfall, vilket kan ge en högre statistisk styrka i studien. Man bör dock vara försiktig med kompositmått. 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 till och med maskera en negativ effekt av behandlingen på viktiga utfall såsom död och hjärt- och kärlhändelser [5].

Man bör också undersöka om det finns så kallade noggrant utvecklade prioriterade utfall (eng. Core Outcome Sets, COS) som kan användas [6]. Prioriterade utfall är framtagna i konsensusprocesser där patienter/brukare, kliniker och forskare bidrar. Syftet är att få en enhetligare rapportering av utfall och att måtten man har använt sig av för att mäta dessa utfall är de mest relevanta för intressentgrupperna. Sammanställningar av framtagna prioriterade utfall finns här.

3.1.4.1 Tröskelvärden

Det finns flera sätt att uttrycka vilken effekt en intervention har. Syftet kan till exempel vara att utvärdera om interventionen har större effekt än jämförelsen, eller om den är likvärdig. När syftet är att mäta om utfallet är likvärdigt behöver man definiera tröskelvärden. Hur stora kan avvikelserna vara för att effekten fortfarande ska anses vara likvärdig? Detta ska anges i projektplanen.

3.1.4.2 Minsta betydelsefulla skillnad

Man kan också vilja relatera effektstorleken till om den är tillräckligt stor för att vara värdefull för den som får interventionen. Om det finns en tillförlitlig bedömning av den så kallade minsta (kliniskt) betydelsefulla skillnaden (eng. Minimal Important Difference, MID, eller Minimal Clinical Important Difference, MCID) kan det vara intressant att relatera en beräknad effekt till denna. Ska MID eller MCID användas i analyserna ska detta anges i projektplanen. Man måste vara medveten om att värdena för MID och MCID kan vara framtagna för delvis andra populationer, och därmed inte helt giltiga för forskningsfrågan. Det kan också kräva en del arbete för att söka och värdera hur tillförlitliga studierna är.

3.1.5 Andra urvalskriterier

PICO behöver ofta kompletteras med andra kriterier för att avgränsa frågan.

3.1.5.1 Setting

Här definieras i vilken miljö som interventionerna ska ges. Översikten kan till exempel vara avgränsad till att interventionen ska ges inom primärvården, kriminalvården, elevhälsan eller särskilda boenden.

3.1.5.2 Studiedesign

De gängse studietyperna för effekter av en intervention är kontrollerade studier, med eller utan randomisering. Randomiserade studier är dock inte lämpliga för att upptäcka ovanligare biverkningar eller komplikationer. Här krävs stora antal deltagare och därför är prospektiva kohortstudier ofta det lämpligaste studieupplägget.

Studier utan jämförelsegrupp accepteras endast i undantagsfall i SBU:s systematiska översikter. För vissa frågor, till exempel effekter av ändrade policies

eller lagstiftning, kan tidsserier (eng. *interrupted time series*, ITS) vara en rimlig design [7] [8] [9]. Här undersöks populationen vid upprepade tillfällen före och efter en introducerad förändring. För att minska risken för att en uppmätt förändring beror på till exempel underliggande samhällstrender snarare än på interventionen måste det finnas flera mätpunkter såväl före förändringen som efter [7].

3.1.5.3 För registerstudier

Om interventionen undersöks i register- eller journalstudier är det viktigt att registret är tillförlitligt. Vilken täckning har registret? Vem har lagt in uppgifter? Verkar data vara rimliga? Är uppgifterna/registret så gamla att det är risk för att de inte är relevanta längre?

3.1.6 Exempel

Nedan finns ett exempel på hur man kan bygga upp ett PICO (Faktaruta 3.2). Exemplet är hämtat från SBU:s rapport om program för att förebygga psykisk ohälsa hos skolbarn [10].

Faktaruta 3.2 Exempel på konstruktion av PICO.

Fråga: Finns det några program som kan förebygga utagerande beteende hos barn i skolåldern?

Frågan omvandlades till ett **PICO**:

P: Barn i åldrarna 2–19 år utan psykiatrisk diagnos. Åldersspannet bestämdes av att studier på förskolebarn skulle inkluderas.

I: Manualbaserade program med primärt syfte att förebygga psykisk ohälsa. Program för att förebygga till exempel drogmissbruk eller mobbning exkluderades därmed. Programmen kunde ges på olika arenor och riktas till barnen eller deras föräldrar.

C: Inga interventioner, andra program.

O: Utagerande beteende mätt med validerade skattningsskalor, psykiatriska diagnoser. Studier som undersökte effekter på föräldrarna exkluderades därmed.

Uppföljningstid: Minst sex månader efter avslutat program.

3.2 Format för sambandsstudier

Kommer senare.

3.3 Format för frågor om diagnostiska test eller bedömningar

3.3.1 Frågan

Diagnostiska test och bedömningsformulär (fortsättningsvis förkortat till test) syftar till att särskilja dem som har ett visst tillstånd, till exempel en sjukdom eller ett socialt problem från dem som inte har det. Tester har ett flertal användningsområden. De kan till exempel användas som stöd för att ställa en diagnos, för att klassificera svårighetsgrad av tillståndet eller för att förutsäga

risken för ett tillstånd i framtiden (prediktion eller prognos). Detta avsnitt fokuserar på frågor om diagnos men tar även upp vissa aspekter som är relevanta för frågor om prediktion.

Utvärderingar av testmetoder handlar oftast om i vilken utsträckning testet klassificerar korrekt (diagnostisk tillförlitlighet, eng. diagnostic accuracy). Men ur patientens/brukarens, klinikerns och samhällets synvinkel är förmågan att klassificera ett tillstånd att betrakta som ett surrogatmått för värdet av testet. Det är mer relevant att undersöka om testet påverkar beslutsprocesser, givna interventioner och hälsa. Detta beskrivs i en modell av Fryback och Thornsbury [11] som delar upp kunskapen om ett diagnostiskt test i sex nivåer från ren teknisk prestanda till testets betydelse på samhällsnivå (Faktaruta 3.3). Frågor på nivåerna 3 till 6, som handlar om konsekvenser av att genomföra ett test, besvaras bäst med randomiserade studier och hälsoekonomiska studier eller modeller.

Den för den enskilde patienten eller brukaren värdefullaste kunskapen är om testet bidrar till ett förbättrat utfall, i form av till exempel lägre dödlighet, lindrigare grad av problem eller bättre livskvalitet, det vill säga nivå 5. Observera att en förutsättning för att ett test är värdefullt på en högre nivå är att det även är värdefullt på de lägre nivåerna – men att det motsatta inte gäller. En hög diagnostisk tillförlitlighet medför alltså inte automatiskt att testet fyller en för patienten viktig funktion.

Faktaruta 3.3 Frågor om värdet av diagnostiska tester [11].

Nivå 1: Teknisk prestanda för testet. Prestanda kan påverkas av så kallade artefakter i omgivningen och användarens skicklighet.

Nivå 2: Diagnostisk tillförlitlighet, dvs. i vilken utsträckning ett test klassificerar tillståndet korrekt, jämfört med en referensmetod. Tillförlitligheten påverkas av den person som tolkar resultatet.

Nivå 3: I vilken utsträckning testet ändrar klinikerns diagnos (processmått)

Nivå 4: I vilken utsträckning testet påverkar vilken åtgärd eller behandling som sätts in (processmått)

Nivå 5: I vilken utsträckning testet påverkar patientens utfall

Nivå 6: I vilken utsträckning testet är kostnadseffektivt

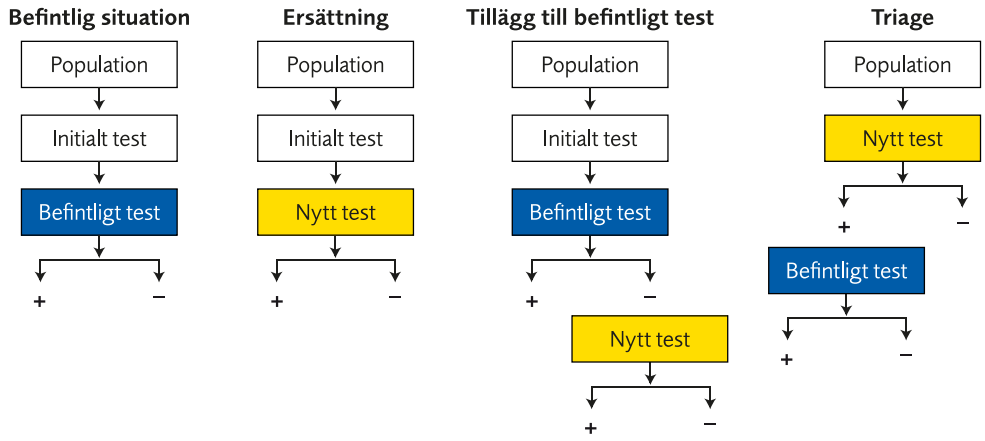
I en systematisk översikt är det önskvärt att kunna utvärdera testets påverkan på patient eller brukare, eller åtminstone att kunna utvärdera om testet leder till förändringar i vårdkedjan (eller motsvarande). I praktiken finns det sällan sådana studier utan översikten kommer sannolikt oftast att begränsas till den diagnostiska tillförlitligheten.

3.3.2 Testets plats i diagnostiken

Det är viktigt att definiera vilken roll i diagnostiken ett test ska spela (se Figur 3.1). Det kan exempelvis ersätta ett annat test som kanske är dyrare, mer komplicerat eller har sämre tekniska prestanda. Det kan också läggas till som ett

extra steg före ett befintligt test (triage). Endast personer som testar positivt på det nya testet går vidare till nästa steg. Triagetestet syftar inte till att öka den diagnostiska tillförlitligheten utan att minska användningen av ett befintligt test som kanske är dyrt eller smärtsamt. Ett tredje alternativ är att testet läggs till efter det befintliga testet (eng. add-on). Här är syftet att öka den diagnostiska tillförlitligheten för en mindre grupp deltagare. Deltagarna i studier om add-on tester kommer att ha annorlunda karakteristika än i studier som undersöker ersättningstest eller triagetest.

Figur 3.1 Testets roll och plats i diagnostiken [12].



3.3.3 Den strukturerade frågan

Frågan om diagnostisk tillförlitlighet besvaras ofta med en tvärsnittsstudie där deltagarna undersöks med både ett experimentellt test och ett referenstest. Den strukturerade frågan har formatet PIRO (Population, Indextest, Referenstest, Utfall).

3.3.4 Population

Här görs samma överväganden som för frågor om effekt av interventioner (se avsnitt 3.1.1).

3.3.5 Indextest

Diagnostisk tillförlitlighet anger hur väl testet som ska utvärderas, indextestet, kan skilja mellan två olika tillstånd, till exempel sjuk och frisk, jämfört med en referensmetod. Indextestet är det test som ska utvärderas och det kan behöva specificeras detaljerat. I vissa fall kan det vara nödvändigt att avgränsa till en viss version av till exempel en medicinsk utrustning.

För en del test finns så kallade tröskelvärden (eng. threshold eller cut-off) som definierar gränsen. Ett problem är att tröskelvärdet kan vara olika för olika populationer. Ett exempel är självskattning av depression med formuläret PHQ-9. Formuläret har ett ursprungligt definierat tröskelvärde men senare studier har visat att optimala tröskelvärden kan vara såväl högre som lägre beroende på om

den som testas har andra samtidiga sjukdomar [13]. Om frågan avser ett visst tröskelvärde ska det specificeras.

Andra test, som till exempel bilddiagnostik, har inget tröskelvärde utan klassificeringen påverkas av faktorer som till exempel skicklighet och erfarenhet hos den som tolkar resultaten, något man bör vara medveten om vid granskningen av studier.

3.3.6 Referensmetod

Referensmetoden kallades tidigare guldstandard (eng. gold standard). Den förutsätts alltid klassificera tillståndet eller problemet korrekt. Det är dock ytterst sällan en referensmetod är perfekt och därför har man lämnat begreppet guldstandard. I många fall finns det inte heller någon etablerad referensmetod. I den situationen får man välja mellan olika typer av konstruerade referensmetoder [14] [15] och de beskrivs kortfattat i Faktaruta 3.4.

Faktaruta 3.4 Alternativa referensmetoder när en etablerad referensmetod saknas [14].

Sammansatt referensstandard: Här kombineras flera tester som var för sig är bristfälliga till ett sammansatt mått.

Panel-eller konsensusdiagnos: Här kombineras resultat av olika tester och undersökningar med kliniska karakteristika och prognostisk information. Valideringen bygger på en stor mängd empiriska data och bestäms ofta genom internationella möten med experter som uppnår konsensus. Ett exempel är DSM-klassifikationen av psykiatriska tillstånd.

Statistiska modeller: Här kombineras klinisk information och andra testresultat i statistiska modeller som genererar en sannolikhet för att en diagnos föreligger.

Validering i efterhand: Här används en prospektiv studiedesign där det faktiska utfallet vid uppföljning relateras till testresultatet. Validering är standardmetod för frågor om prediktion.

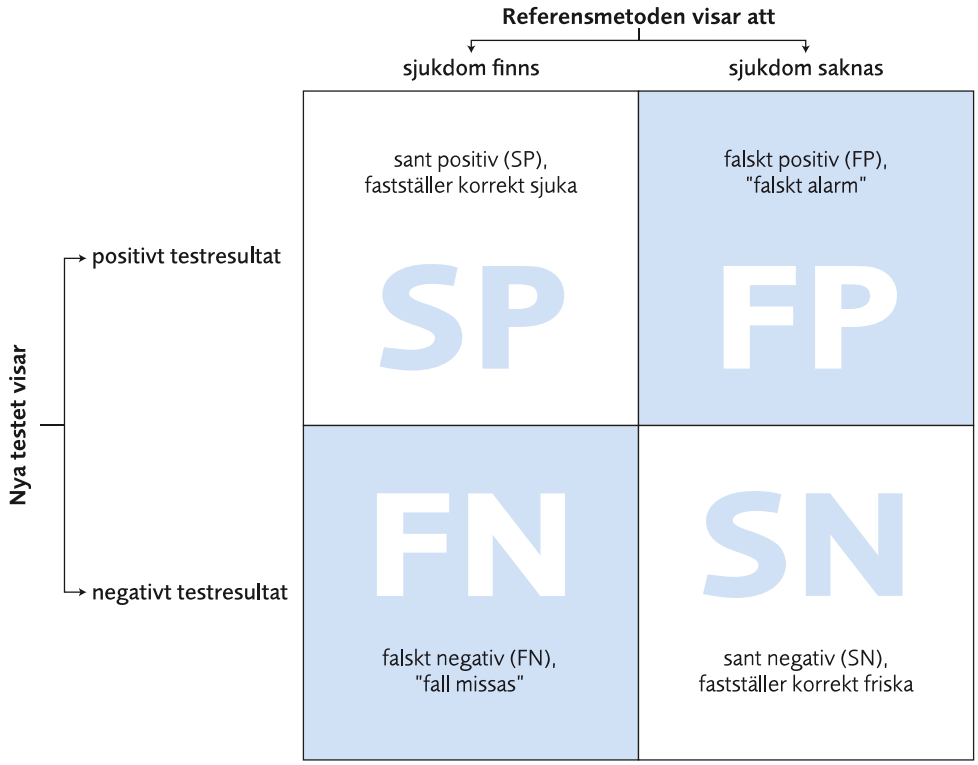
3.3.7 Utfall

3.3.7.1 Vid ett tröskelvärde

Utfallet i en diagnostisk tillförlitlighetsstudie är andelar av deltagarna som är sant respektive falskt klassificerade. Det ger fyra utfall: sant positivt, falskt positiv (dvs. positiv enligt indextestet men inte referenstestet), sant negativ och falskt negativ (dvs. negativ enligt indextestet men inte referenstestet). Utfallen kan läggas in i en så kallad fyrfältstabell, se Figur 3.2.

Utifrån fyrfältstabellen kan man beräkna indextestets sensitivitet (känslighet) respektive specificitet (träffsäkerhet). Sensitivitet och specificitet har ansetts vara relativt okänsliga för tillståndets prevalens, men en systematisk översikt har kommit fram till att prevalensen kan ha en betydande påverkan [16].

Figur 3.2 Fyrfältstabell för diagnostisk tillförlitlighet. Sensitiviteten definieras som andelen av dem som har tillståndet ifråga som testar positivt, medan specificiteten definieras som andelen av dem som inte har tillståndet ifråga som testar negativt.



Definition:

Sensitivitet = sannolikheten för positivt testresultat när man har sjukdomen

Specificitet = sannolikheten för negativt testresultat när man är frisk

Formler:

$$\text{Sensitivitet (\%)} = \frac{SP}{(SP + FN)} \times 100$$

$$\text{Specificitet (\%)} = \frac{SN}{(FP + SN)} \times 100$$

Sensitivitet och specificitet är inbördes beroende av varandra och vanligen leder en ökning av sensitivitet till en sänkning av specificiteten. I praktiken är ofta ett av måtten viktigare än det andra. Ibland är en hög sensitivitet avgörande, exempelvis då man eftersträvar att fånga in så många som möjligt med ett specifikt tillstånd. Detta innebär dock en samtidig ökad risk för falskt positiva resultat. Möjliga konsekvenser av falskt positiva resultat är onödig oro och att utsättas för en intervention som inte behövs och som i sig kan medföra risker. I andra fall kan det vara viktigare med en hög specificitet, något som ökar risken för falskt negativa resultat. Möjliga konsekvenser för patienten eller klienten är en fördröjd diagnos och att tillståndet försämras. Det bör finnas ett resonemang i projektplanen om vilka mått som är de väsentligaste.

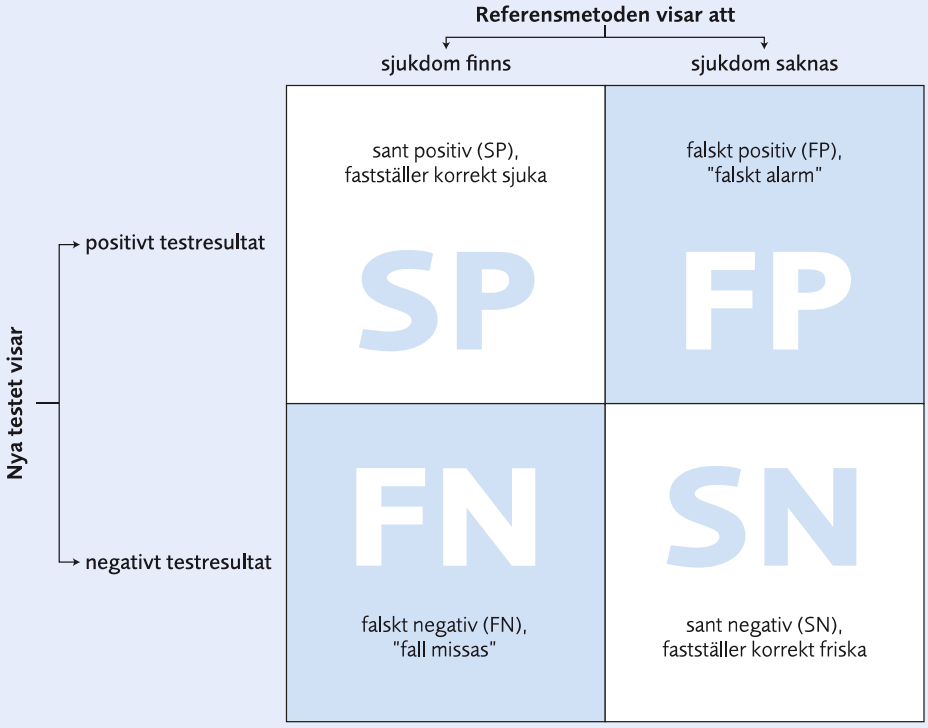
Ibland kan det vara mer värdefullt att få en uppfattning om de så kallade positiva och negativa prediktionsvärdena, PPV respektive NPV. PPV är ett mått på sannolikheten att en person med ett positivt testresultat verkligen har tillståndet,

det vill säga SP/(SP+FP). PPV (och NPV) kommer därför att vara beroende av prevalensen av det aktuella tillståndet i den studerade populationen. Ju lägre prevalensen är desto lägre kommer också PPV vara.

Från fyrfältstabellen kan man även beräkna den diagnostiska oddskvoten, DOR (eng. diagnostic odds ratio) och positiva, respektive negativa likelihood ratios (LR). Läs mer om dessa nedan.

Den diagnostiska oddskvoten (DOR) samt positiva och negativa likelihood ratios (LR)

Figur 3.3 Fyrfältstabell för diagnostisk tillförlitlighet. Sensitiviteten definieras som andelen av dem som har tillståndet ifråga som testar positivt, medan specificiteten definieras som andelen av dem som inte har tillståndet ifråga som testar negativt.



Definition:

Sensitivitet = sannolikheten för positivt testresultat när man har sjukdomen

Specificitet = sannolikheten för negativt testresultat när man är frisk

Formler:

$$\text{Sensitivitet (\%)} = \frac{SP}{(SP + FN)} \times 100$$

$$\text{Specificitet (\%)} = \frac{SN}{(FP + SN)} \times 100$$

Från fyrfältstabellen (Figur 3.3) kan man även beräkna den diagnostiska oddskvoten, DOR (eng. diagnostic odds ratio) och positiva, respektive negativa likelihood ratios (LR) (Faktaruta 3.5). Nackdelen med DOR är att den ger mindre information och har begränsad användning för systematiska översikter.

Faktaruta 3.5 Definition och beräkning av likelihood ratio och diagnostisk oddskvot utifrån ett fiktivt exempel.

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.

3.3.7.2 När det finns flera tröskelvärden

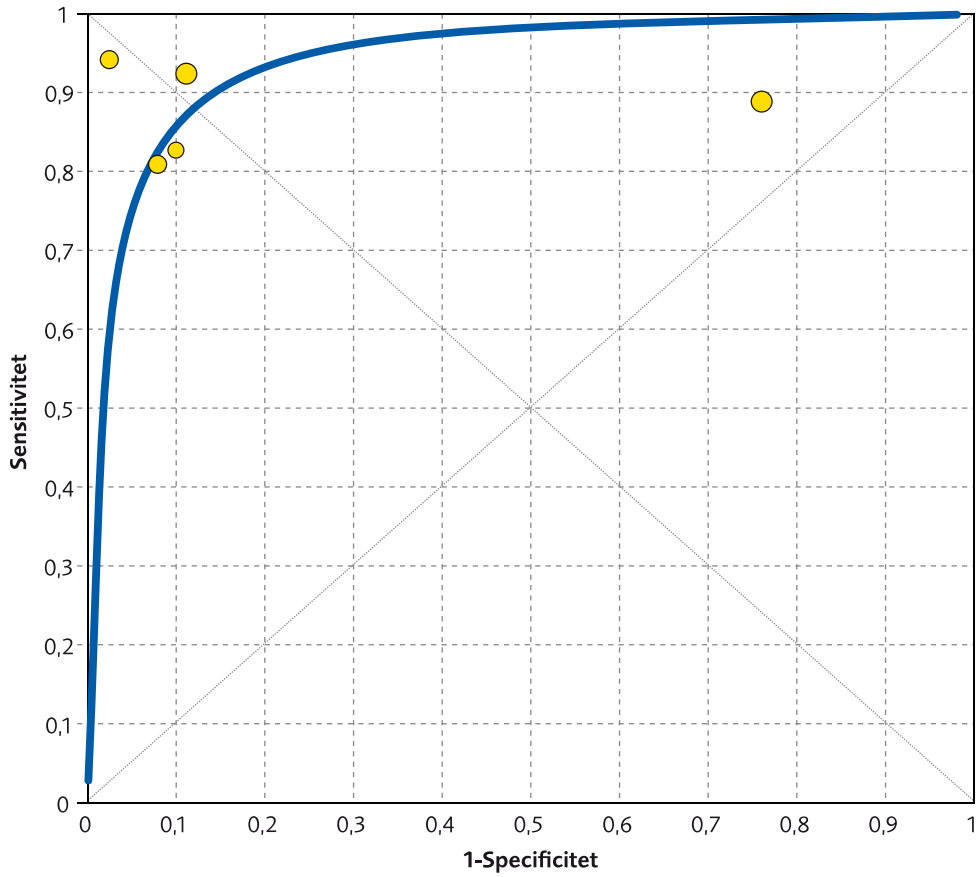
Vissa test är dikotoma, det vill säga de visar endera att tillståndet finns eller att det inte finns. Men om testresultaten uttrycks som poäng på en skattningsskala eller är ett kontinuerligt mått som till exempel blodtryck måste det finnas ett tröskelvärde som definierar gränsen mellan att "ha", eller "inte ha" tillståndet. Resultaten omvandlas sedan till om de ligger över eller under tröskelvärdet, det vill säga "ja" eller "nej". Ibland har testet utvärderats för flera tröskelvärden. Om ett tröskelvärde är mer etablerat än andra kan man nöja sig med att beräkna sensitivitet och specificitet vid det tröskelvärdet. För andra tester, till exempel radiologiska bilder, saknas ett etablerat tröskelvärde.

För att få en uppfattning om hur ett test fungerar vid olika tröskelvärden kan man konstruera en så kallad Receiver Operating Characteristics curve, (ROC-kurva [17] [18]). ROC-kurvan visar andelen sant positiva (y-axeln) mot andelen falskt positiva (x-axeln), det vill säga sensitivitet mot 1–specificitet för varje tänkbart tröskelvärde (Figur 3.4).

ROC-kurvans utseende beror på karakteristika för patienterna/brukarna och vilket ”spektrum” av tillståndet som är representerat i studien, till exempel olika svårighetsgrader.

Ytan under ROC-kurvan (eng. area under the curve, AUC) är ett globalt mått som summerar den diagnostiska tillförlitligheten över alla tröskelvärden. AUC kan ha ett värde mellan 0 och 1, där 0 indikerar ett helt otillförlitligt test och 1 indikerar ett perfekt test. Ett AUC på 0,5 tyder på att testet inte kan skilja på dem som har tillståndet och dem som inte har det, det vill säga att slumpen avgör, och representeras av en diagonal linje mellan punkterna (0,0) och (1,1).

Figur 3.4 Exempel på Receiver Operating Characteristics curve (ROC) med tillhörande area under kurvan (AUC) som summerar fem studier (gula punkter) [19].



AUC = Area under kurvan; **Q*-index** = Den punkt där sensitivitet och specificitet är lika stora, vilket är punkten närmast övre vänstra hörnet på kurvan; **SE** = Standardfel

AUC täcker hela området från 0 till 100 procent specificitet respektive 0 till 100 procent sensitivitet. I praktiken är man ofta intresserad av en mer begränsad del av ROC-kurvan, till exempel området med högst specificitet, vilket medför att AUC-värdet kan bli missvisande. AUC som mått på diagnostisk tillförlitlighet har kritiserats, till exempel i artikeln av Halligan och medarbetare [20].

Ett problem är att det inte finns några etablerade gränser för när AUC kan anses vara tillräckligt hög. Gränser för låg, måttlig och hög diagnostisk tillförlitlighet måste därför beslutas och definieras i projektplanen. Två SBU-rapporter har använt följande gränser: 0,70–0,80 (modest); 0,81–0,90 (måttlig) och >0,90 för hög diagnostisk tillförlitlighet [21] [22]. I ett annat exempel, vid diagnostik av sepsis, sattes gränserna till: 0,6–0,7 (låg); 0,7–0,9 (måttlig) och >0,9 för hög diagnostisk tillförlitlighet [23].

3.4 Format för frågor om upplevelser, erfarenheter och värderingar

Frågor om personers upplevelser, erfarenheter och värderingar av en viss företeelse, ett så kallat fenomen, har blivit alltmer intressant som en del i utvärderingar [24]. Frågorna besvaras bäst med hjälp av metoder som intervjuer eller observationer, det vill säga forskning med kvalitativ ansats. Även enkätstudier kan undersöka uppfattningar och erfarenheter och enstaka

syntesmetoder accepterar resultat från såväl kvantitativ som kvalitativ forskning [25].

Kvalitativ forskning bottnar i olika traditioner som till exempel filosofi, antropologi och sociologi, som i sin tur blivit basen för olika forskningsansatser [26]. Några vanliga exempel är fenomenologi och hermeneutik som utgår från filosofiska teorier och grounded theory som utgår från sociologi. Valet av ansats bestäms av studiens syfte och relation till teori, där syftet kan vara att till exempel generera ny teori, att testa befintliga teorier eller vara tillämpat, exempelvis som en del i en utvärdering av komplexa metoder [26]. Ansatsen kommer i sin tur att påverka valet av metoder för att samla in, analysera och tolka data. En sammanställning av exempel på olika forskningsansatser hittar du nedan.

Tabell 3.1 Exempel på kvalitativa forskningsansatser.

Ansats	Beskrivning	Exempel på studie inom hälso- och sjukvården	Exempel på studie inom socialtjänst
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" [27]	Djupintervjuer gjordes med 28 hemlösa ungdomar i åldern 18 till 24 år, varav 20 var män och 8 var kvinnor Syftet var att studera mångfalden i familje-miljöegenskaper och berättelser om missbruk bland hemlösa unga vuxna Studiens fynd kan delas in enligt följande: "misshandel inom familjen", "misshandel av vårdnadsgivaren", "att bli avvisad" och "avståndstagande vårdnadsgivare" [28]
Fenome-nologi	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 [29]	Syftet med studien var att utforska och beskriva erfarenheter hos sjuksköterskor av dödsfall vid ett lång-tidsboende för äldre Semi-strukturerade intervjuer hölls med sju sjuksköterskor. Författarna fann följande huvudteman: "livets sista resa", "familj" och "professionell vårdgivare" [30]

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 [29]	Denna empiriska studie utforskar och beskriver variationen i hur evidensbaserad praktik uppfattas inom socialtjänst. Fjorton semi-strukturerade intervjuer utfördes med politiker, chefer och verksamhetsledare vid tre socialtjänstkontor i Sverige De huvudsakliga fynden om hur evidensbaserad praktik uppfattas är följande kategorier: 1) fragmenterat, 2) diskursivt, 3) instrumentellt, 4) mångfacetterat och 5) kritiskt [31]
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 [32]	Med hjälp av ostrukturerade intervjuer undersöktes hur tio fosterföräldrar från landsbygden i nordöstra USA upplevde att det var att ta hand om fosterbarn med speciella hälso- och sjukvårdsbehov Följande teman identifierades: 1. tillgänglighet till hälso- och sjukvård, 2. överväldigad och oförberedd, 3. beslutsfattande, 4. känna sig isolerad samt 5. känna sig stigmatiserad [33]
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 [34]	Syftet med studien var att utforska vardagsarbetet, som utförs av så kallade "case managers" med fokus på deras erfarenheter. Datainsamlingen gjordes med hjälp av deltagande observation med fältanteckningar, samt gruppintervjuer och individuella intervjuer med nio "case managers" Ett övergripande tema identifierades: rådande yrkesidentitet utmanas. Det fanns tre underteman: 1) anpassa sig till bekant arbete i en obekant roll, 2) strävan efter att förbättra ett hälsosystem genom en ny roll, 3) förtroende är vitalt för att kunna

			representera den äldre [35]
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 [29]	Syftet med denna norska studie var att få en ökad förståelse för typ och kvalitet i förhållandena mellan personer med demens, vårdgivare i hemmet och yrkesverksamma vårdare, samt hur dessa förhållanden påverkar personer med demens och deras möjlighet att vara person i betydelsen självständig person med individuella mänskliga egenskaper Studien baseradespå tio fall. En semi-strukturerad intervjuguide användes vid intervjuer med familjemedlemmar som agerade vårdare i hemmet samt med de yrkesverksamma vårdarna. Fältanteckningar togs efter deltagandeobservation av interaktionen mellan personer med demens och yrkesverksamma vårdgivare under morgonrutiner eller aktiviteter på ett dagvårdscenter [36]
Kvalitativ innehålls-analys	Innehållsanalys innebär vanligen att forskaren genom upprepad läsning av en text identifierar meningsenheter 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 innehållsanalys används för att identifiera likheter och skillnader i en text ansågs ansatsen lämplig för ändamålet [29]	Kvalitativa forskningsmetoder användes för att utforska olika intressenters uppfattningar av olika influenser som påverkar fosterbarns hälsa. Semi-strukturerade intervjuer genomfördes i fokusgrupper bestående av fosterbarn, fosterföräldrar och yrkesverksamma inom fosterhemsplacering. Insamlad data analyserades med hjälp av kvalitativ innehållsanalys [37]
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 aktionsforskning är ett deltagarbaserat och interaktivt förhållningssätt. Detta innebär att alla deltagare, både forskare och	Syftet med studien var att arbeta fram och att införa en lämplig modell för handledning av vårdpersonal inom kriminalvården [34]	Författarna sökte svar på bl.a. följande frågor: 1) hur kan man ta i beaktande ungdomars åsikter när man utvecklar praxis inom socialtjänst? 2) vilka förutfattade meningar om ungdomar tar socialtjänstarbetaren med sig i sitt arbete? Studien beskriver ett speciellt tillvägagångssätt – en

	praktiskt verksamma personer arbetar tillsammans		metod – som gör att frågorna kan adresseras med en sammanhållen struktur för aktionsforskning Den grundläggande idén var att locka fram åsikter från ungdomar som har haft erfarenhet av socialtjänsten och presentera dessa åsikter för yrkesverksamma inom socialtjänsten, som sedan ombads att ändra sina rutiner [38]
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 kompletterande 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 [29]	Studien syfte var att utforska effekterna av fosterhemsupplevelsen för fosterfamiljens biologiska barn. Barn från åtta familjer intervjuades och datamaterialet analyserades med hjälp av narrativ analysmetod [39]

Det kan finnas flera syften med en kvalitativ översikt. Ett syfte som förekommer i flera SBU-rapporter är att utforska erfarenheter, upplevelser och uppfattningar av sjukvård eller omsorg, ett annat att utforska upplevelser av att ha ett visst tillstånd, som till exempel i SBU:s rapport om tandförluster [40]. Två andra syften är mera länkade till att utvärdera interventioner. En systematisk översikt om effekter av en intervention kan kompletteras med erfarenheter och upplevelser av att ge respektive få interventionen, vilket kan bidra till en rikare bild av interventionens effekter. En variant är när en kvalitativ översikt utforskar hinder och underlättande faktorer för att implementera en intervention [41]. I Cochranes handbok för systematiska översikter finns vägledning för hur resultat från kvantitativ och kvalitativ syntes kan läggas ihop [42].

Resultaten av en kvalitativ studie uttrycks oftast som teman eller kategorier. Kvalitativ forskning är hårt bunden till sammanhanget (kontexten), som omfattar såväl den studerade populationen som miljön (eng. setting). Fynden kan även vara påverkade av omgivningsfaktorer som lagstiftning och policies när data samlas in.

Det finns två typer av strukturerade frågor beroende på metod för syntesen [24]. Det vanligaste i en HTA-rapport är en så kallad fixerad fråga som fungerar som PICO. Det mest etablerade formatet är SPICE som består av fem komponenter (Tabell 3.2); Setting (sammanhang), Perspective (perspektiv), Intervention/interest (intervention), Comparison (jämförelse), och Evaluation (utvärdering). Ett mer detaljerat format är PerSPecTIF [42], där kontexten delas upp i Setting och Environment. Formatet lämpar sig väl för frågor om komplexa interventioner och implementering. Ibland kan dock frågan vara

öppen och formuleras färdigt under syntesens gång i analogi med grounded theory.

Tabell 3.2 SPICE-format för strukturering av forskningsfrågan.

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

3.4.1 SPICE

3.4.1.1 Setting

Setting omfattar till exempel geografiskt område, miljö (t.ex. primärvård eller fängelse), lagstiftning och policies som kan påverka fenomenet, och när studierna ska vara genomförda för att vara relevanta.

3.4.1.2 Perspektiv

Denna komponent definierar vems perspektiv som översikten gäller, det vill säga populationen av intresse. I vissa fall kan det vara värdefullt att specificera subgrupper.

3.4.1.3 Intervention/intresse

Denna komponent specificerar fenomenet på den detaljnivå som behövs. Om fenomenet är för översiktligt definierat kommer det att bli svårt att bedöma relevansen i det vetenskapliga underlaget. Även om frågan kan vara avgränsad till att till exempel utforska upplevelser av sjukvården kan ibland sökningen behöva breddas så att man till exempel accepterar studier som handlar om att leva med ett tillstånd.

3.4.1.4 Jämförelse

I de fall där det är aktuellt att ha en jämförelse, till exempel vid frågor om upplevelser och erfarenheter av interventioner, ska jämförelsen beskrivas tillräckligt detaljerat.

3.4.1.5 Utvärdering

Här definieras om frågan gäller resultat i form av till exempel upplevelser, erfarenheter, åsikter eller observerade beteenden.

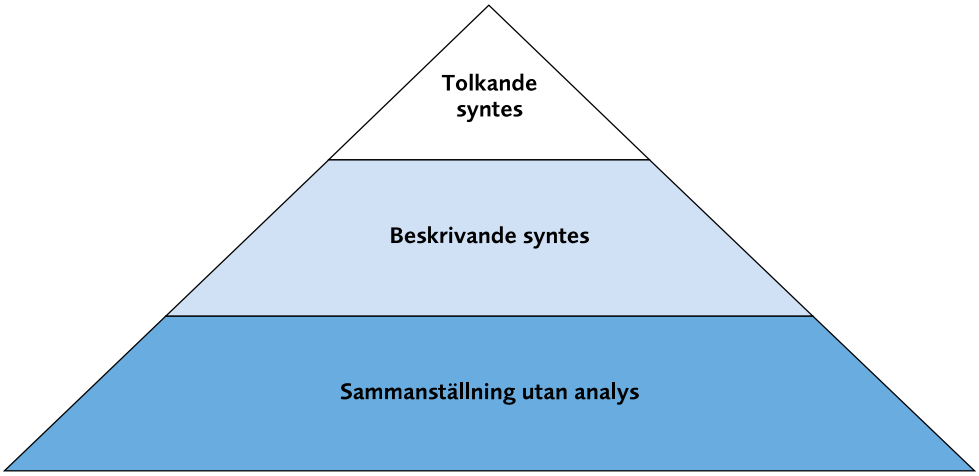
3.4.2 Teori

En teori, modell eller teoretiskt ramverk kan underlätta uppgiften att identifiera viktiga faktorer i den strukturerade frågan. En beteendeteori eller social teori kan till exempel vara ett stöd för att förfina frågan, något som beskrivs mer detaljerat av Noyes och medarbetare [43].

3.4.3 Val av syntesmetod

Den tilltänkta syntesmetoden ska specificeras i projektplanen. Valet av syntesmetod beror främst på forskningsfrågan men även praktiska aspekter såsom tid och tillgänglig expertis spelar in. Grovt sett kan man dela in metoderna i beskrivande eller tolkande, där vissa metoder har såväl beskrivande som tolkande inslag (Figur 3.5). Valet av metod påverkar också litteratursökning och sökstrategi, se Kapitel 4.

Figur 3.5 Indelning av syntesmetoder.

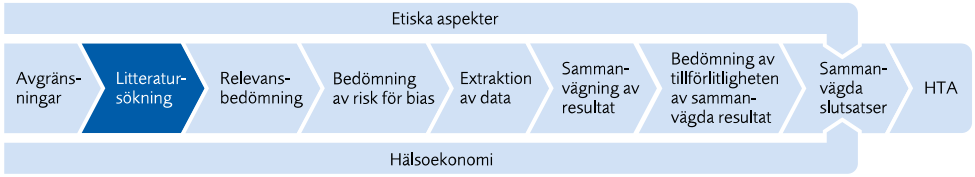


3.4.4 Reflexivitet

Begreppet reflexivitet rör dialogen mellan forskaren (i det här fallet projektgruppen) och forskningen (i det här fallet den systematiska översikten). Den kan vara prospektiv eller retrospektiv. Prospektiv reflexivitet handlar om den påverkan översiktsförfattarna har på översikten. Den omfattar överväganden om hur författarnas förförståelse i form av kunskap, synsätt och uppfattningar kan påverka val av både frågor och metod, men också vilka tolkningar som görs under syntesen. Retrospektiv reflexivitet ger en möjlighet att överväga om projektprocessen och de resultat som framkommer lett till att förförståelsen förändrats.

Projektgruppens förförståelse och dess möjliga påverkan på metodval, liksom strategier för att minska påverkan, ska beskrivas i projektplanen och i rapportens metodavsnitt. Om projektgruppen kommer fram till att påverkan är stor kan det finnas skäl till att engagera ytterligare författare med andra perspektiv. Förförståelsen, och om den har ändrats under projektets gång, bör även tas upp i rapportens diskussion om resultaten.

4. Litteratursökning



I Kapitel 2 beskrivs riktlinjer för hur arbetet med systematiska översikter ska utformas och dokumenteras (PRISMA:s checklista) [44]. Avsnitt 7 och 8 i denna checklista ger anvisningar om hur litteratursökningen ska rapporteras. För att upprätthålla kraven på översiktens transparens och reproducerbarhet ska alla källor som har använts vid sökningen anges och beskrivas. Det är också viktigt att ange tidpunkten för senaste sökning, eftersom denna ofta skiljer sig betydligt från när översikten publicerats. En fullständig, reproducerbar dokumentation av sökstrategin för åtminstone en databas ska också bifogas. Dokumentationen gör det möjligt att se om arbetet med litteratursökningen följer internationell standard. Förutom SBU:s metodbok och internationella metodböcker [45] [46] [47] [48] [49] som ger utförliga anvisningar om hur litteratursökningen ska utformas, forskas det aktivt inom området och en omfattande mängd av vetenskapligt granskade metodartiklar ges också ut. Mer övergripande information om sökning för systematiska översikter hittas bland annat i dessa publikationer av Atkinson och medarbetare [50] samt Cooper och medarbetare [51]. Webbplatsen SuRe info (Summarized Research in Information Retrieval for HTA) som är en del av HTAi Vortal, är också en viktig källa för att följa den internationella metodutvecklingen.

Det här kapitlet handlar om litteratursökningen som en del av SBU:s projektprocess, med fokus på sökning efter vetenskapliga originalartiklar i internationella ämnesdatabaser. Även kompletterande söksätt och sökning av grå litteratur tas upp.

4.1 Litteratursökningen – en del av projektprocessen

Arbetet med att ta fram en så heltäckande sökstrategi som möjligt är ett samarbete mellan informationsspecialist, projektledare och projektets sakkunniga. Det är en stor fördel att involvera informationsspecialisten redan i samband med att projektplanen utformas, eftersom informationsspecialistens arbete med sökstrategin effektiviseras genom en ökad förståelse för frågans olika aspekter. Samtidigt kan informationsspecialistens kunskap och erfarenheter av att omsätta en fråga till en sökstrategi bidra till att strukturera frågan. Studier har visat att när informationsspecialisten deltar i projekten ökar kvalitén på litteratursökningen, framför allt genom att sökningen blir reproducerbar i enlighet med PRISMA-statement [52] [53].

Utgångspunkten för litteratursökningen är alltid utvärderingens frågeställning, som struktureras i projektplanen. Sökningen görs i flera steg: förberedande

sökningar, testsökning och huvudsökning. Innan huvudsökningarna påbörjas ska projektplanen vara fastställd och godkänd. I slutet av projektet görs en uppdateringssökning så att underlaget är så aktuellt som möjligt.

4.1.1 Före projektstart: identifiera redan gjorda översikter

Innan ett projekt startar bör man kontrollera om liknande projekt pågår i någon annan HTA-organisation eller om det finns andra aktuella systematiska översikter som kan besvara frågan. Oavsett om syftet med projektet är att identifiera så många som möjligt av de relevanta systematiska översikter som publicerats inom ämnet, eller om syftet är att identifiera originalartiklar för att sammanställa en systematisk översikt, gäller att flera databaser måste sökas. De befintliga stora internationella ämnesdatabaserna behöver kompletteras med ett antal specialdatabaser och utvalda organisationers webbplatser.

Viktiga databaser är exempelvis

- Cochrane Library
- Epistemonikos
- Evidence search (NICE)
- International HTA Database

Databaser inom det sociala området som bör kontrolleras är (exempel):

- Social Care Online
- Campbell Collaboration

SBU registrerar alltid sina pågående utvärderingsprojekt och kartläggningar, oavsett ämnesområde, i den fritt tillgängliga databasen PROSPERO. SBU är medlem i det europeiska nätverket EUnetHTA[1] och registrerar därför också pågående utvärderingar i den interna medlemsdatabasen POP database.

1. EUnetHTA (European Network for Health Technology Assessment) stödjer arbetet med HTA-rapporter i Europa.

Exempel på databaser/webbplatser som innehåller systematiska översikter och HTA-rapporter

Agency for Healthcare Research and Quality (AHRQ)
HTA-organisation (USA)

Canadian Agency for Drugs and Technologies in Health (CADTH)
Nationell HTA-organisation (Kanada)

Cochrane Library
Innehåller flera deldatabaser, bland andra Cochrane Database of Systematic Reviews

Campbell Library
Innehåller systematiska översikter inom ämnesområdena socialt arbete, kriminologi, utbildning

Epistemonikos
Innehåller systematiska översikter inom hälso- och sjukvård samt översikternas inkluderade artiklar

Evidence Search
Innehåller bland annat systematiska översikter och andra typer av sammanställd kunskap, både från den egna organisationen (National Institute for Health and Care Excellence, NICE, Storbritannien) och från externa källor,

Folkehelseinstituttet
Nationell institution som bland annat publicerar systematiska översikter/HTA-rapporter inom hälso och sjukvård samt socialt arbete (Norge),

International HTA database
Innehåller mer än 16 000 rapporter från över 120 olika HTA-verksamheter i hela världen.

KSR Evidence
Innehåller systematiska översikter inom hälso- och sjukvård som utvärderats med ett verktyg som baseras på ROBIS. Licensierad. Kleijnen Systematic Reviews Ltd (KSR)

SBU, Statens beredning för medicinsk och social utvärdering
Nationella och regionala HTA-rapporter (Sverige)

Social Care Online
Innehåller bland annat systematiska översikter, myndighetspublikationer, och originalstudier från Storbritannien inom ämnet socialt arbete.

4.1.2 Testsökning

Inför projektstart formulerar informationsspecialisten, i samarbete med projektledaren, sökstrategier för testsökningar. Testsökningarna syftar till att klarlägga bland annat:

- Hur relevanta studier är indexerade och vilka termer som förekommer i titel och abstrakt
- Om frågorna är tillräckligt väldefinierade, eller om de behöver förtydligas
- Sökmängder som kan förväntas

Projektets sakkunniga har en viktig roll i samband med testsökningarna. De kan förse informationsspecialisten med centrala artiklar och översikter som är relevanta för frågeställningen, och som kan användas för att utveckla sökstrategierna. Vid testsökningen kontrollerar informationsspecialisten vilka ord och fraser som generellt används i abstrakt och titlar, författarnas egna ämnesord, vilka kontrollerade ämnesord som används samt om dessa centrala artiklar, och även de inkluderade artiklarna i översikterna, verkligen fångas av sökningen. Sakkunniga kan också bidra med begrepp och uttryck från sina respektive

forskningsområden och bedöma om sökresultatet är passande för projektets fråga(or) eller om sökstrategin behöver korrigeras.

Informationsspecialisten och sakkunniga kan samarbeta på olika, och ibland kompletterande sätt. Sökstrategierna och sökresultatet kan diskuteras exempelvis via e-post, eller genom fysiska såväl som online-möten. De sakkunniga kan också få möjlighet att bekanta sig med det preliminära sökresultatet, till exempel genom Collections från PubMed, eller i form av ett bibliotek från ett referenshanteringssystem, för att därefter lämna synpunkter till informationsspecialisten. Projektledarens roll kan variera, men det är viktigt att denne är väl insatt i hur arbetet fortskrider.

4.1.3 Att skapa sökstrategier

Som beskrivits tidigare (se Kapitel 3) 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.

Nedan hittar du mer information om A) hur man utformar en sökstrategi; från PICO till sökning, hur man skapar sökblock, om booleska operatorer, parentessökning, olika typer av sökord, hur man identifierar söktermer, hur man kan avgränsa, litteratursökningens omfattning – en balansgång, B) litteratursökning i hälsoekonomiska utvärderingar, C) att söka opublicerade data och övrig grå litteratur samt D–E) om att söka studier med kvalitativ ansats.

A) Hur man utformar en sökstrategi

Från PICO till sökning

Som en hjälp i arbetet med att strukturera frågeställningen använder man för interventionsstudier förkortningen PICO (population, intervention, comparison/control, outcome). Ibland kan man även se PICOS, där S står för study design. För diagnostik används PIRO (population, index test, reference standard, outcome) och för studier som bygger på kvalitativ data kan projektets frågor struktureras med hjälp av SPICE (setting, perspective, intervention, comparison, evaluation). Det finns även andra sätt att strukturera frågeställningarna. För mer information, gå till webbplatsen för SuRe Info (Summarized Research in Information Retrieval for HTA).

Skapa sökblock

När man formulerar en sökstrategi använder man sig vanligtvis av det som på engelska kallas för en "building block strategy", eller det som på svenska kan kallas blocksökning. Ett sökblock är alla tänkbara synonymer/fraser som kan användas för att beteckna till exempel ett sjukdomstillstånd, en insats eller en studiedesign. Ett sökblock består både av indexeringsord, hämtade från ordlistan för den specifika databasen (tesaurus), och av fritextord. För att inte riskera att man missar studier omvandlar man bara några delar av PICO till sökblock. Vanligen använder man sökblock för population och intervention, men ibland lägger man även till ett block med termer för studiedesign. Vissa delar av PICO kan ibland också motsvaras av två block i sökningen. Om frågeställningen till exempel handlar om populationen "äldre personer med urininkontinens" kan detta förslagsvis motsvaras av två block; ett block för äldre personer och ett block för urininkontinens. Först söker man varje block var för sig, och sedan kombinerar man dem med varandra för ett slutgiltigt sökresultat.

Det finns områden där man behöver komplettera blocksökningen med andra metoder för att utforma sökningen, till exempel vid komplexa interventioner och vid utvärdering av diagnostiska

metoder. Exempel på sådana sökningar är sökningar som består flera smala sökstrategier, ofta med olika begränsningar, som man sedan kombinerar [49] [54].

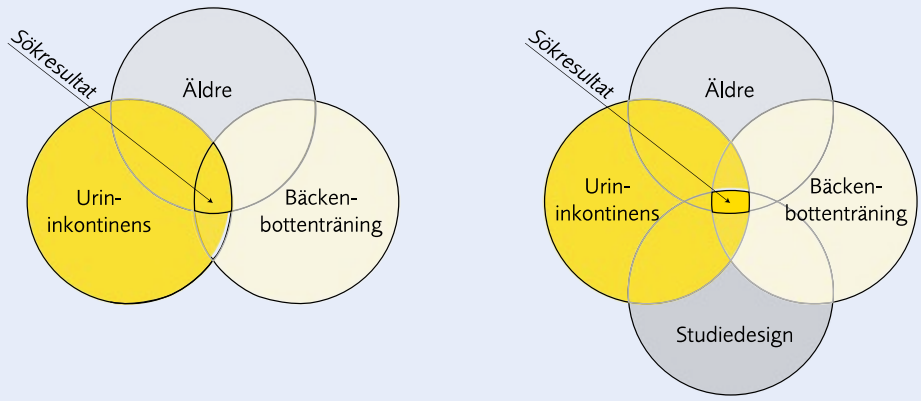
Booleska operatorer och närhetsoperatorer för att kombinera sökord

De enskilda blocken av söktermer som ska ingå i sökningen skapar man genom att kombinera uttryck och termer med en boolesk operator. De flesta internationella databaser, dock inte PubMed, erbjuder också möjligheten att söka med så kallade närhetsoperatorer. Det ser lite olika ut hur dessa operatorer skrivs men information finns i respektive databas hjälpsida. Beroende på vilken typ av närhetsoperator man använder sig av kan man styra i vilken ordning söktermerna får stå samt hur många ord som får stå mellan de angivna termerna. De booleska operatorerna "AND", "OR", "NOT" ger databasen specifika instruktioner, och ska inte blandas samman med ordens vardagliga betydelse. Inom varje block av söktermer kombinerar man 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 den ena eller den andra söktermen eller alla termer som förekommer i blocket. Genom att använda "OR" garderar man sig för den mångfald av olika uttryck som kan användas i olika artiklars referenser för en och samma sjukdom, intervention etc. Eftersom man söker på flera olika synonymer eller termer som representerar samma begrepp blir sökresultatet i sökningar med "OR" mellan söktermerna större än om man bara hade sökt med ett specifikt sökord.

När varje block av sökord är sökta kombinerar man dem med ett booleskt "AND" mellan varje block. 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 funktion för sökhistorik som finns i de flesta databaser.

Figur 4.1 Sökresultat med den booleska operatorn "AND".



Den booleska operatorn "NOT" använder man för att ge databasen instruktionen att något inte ska förekomma i sökresultatet. Vanligen används NOT med stor försiktighet då det finns en risk att man missar relevanta referenser. Om man exempelvis är intresserad av typ 2 diabetes men inte typ 1, och begränsar sökresultatet med "NOT" för termer om diabetes typ 1, kan man missa referenser som man vill att ska ingå i sökresultatet, om de till exempel nämner att "de inte utvärderat typ 1.

Parentessökning

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

Exempel:
gambl* AND (excessive OR pathologic* OR addict* OR disorder* OR problem* OR heavy OR sever* OR compulsive)

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

Olika typer av sökord – indexeringsord

En sökstrategi till en systematisk översikt består av både indexeringsord och fritextord, för att man ska 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. MEDLINE:s (PubMed) tesaurus kallas exempelvis för MeSH, PsycINFO:s kallas "Thesaurus of Psychological Index Terms" och Sociological Abstracts tesaurus heter "Thesaurus of Sociological 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, det går alltså inte att överföra samma söktermer rakt av från en databas till en annan.

Huvuddelen av alla artikelreferenser som läggs in i en databas indexeras. Det vill säga att ett antal termer ur en tesaurus läggs till varje referens (referensen "taggas" alltså), antingen av en indexerare eller med hjälp av en automatiserad indexeringsprocess. Dessa indexeringsord ska beskriva innehållet i en artikel och kan ibland även ange studiedesign, publikationstyp med mera. 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.

Olika typer av sökord – fritextord

Den andra typen av sökord som man använder kallas fritextord. Det är söktermer som man väljer för att matcha ord som förekommer i databasens referens till varje specifik artikel. Referensen är uppdelad i olika fält och en vanlig begränsning är att låta fritextorden matcha ord som finns i fälten för titlar, abstrakt och författarnas egna ämnesord.

Fördelar och nackdelar med indexeringsord respektive fritextord

Fördelar med att söka med hjälp av databasernas indexeringsord är att de är enhetliga. Artikelns abstrakt ska idealiskt beskriva en artikels innehåll, men att söka på ord i en beskrivande text kan leda till irrelevanta träffar. Med indexeringsord behöver man inte ta hänsyn till synonymer och stavningsvarianter som man måste göra med fritextord, vilket är en stor fördel. En nackdel kan vara att de ibland är för generella för att passa den aktuella frågeställningen. Artikelförfattarens val av titel och hur abstraktet är skrivet kommer ha betydelse för hur artikeln indexeras, vilket innebär att den mänskliga faktorn vad gäller felindexering också måste beaktas.

Fördelar med fritextord är att man med hjälp av dessa även hittar studier som ännu inte hunnit bli indexerade. Det betyder att för att fånga de allra senaste publicerade artiklarna i till exempel den viktiga databasen MEDLINE/PubMed, räcker det inte att söka med indexeringsord. En kombination av indexeringsord och fritexttermer kommer alltså att behövas. Fritexttermer kan också vara till hjälp när databasens indexeringsord är för generella för att passa den aktuella frågeställningen, exempelvis vid specifika namngivna interventioner.

Identifiera söktermer

När man skapar en sökstrategi identifierar man både indexeringstermer och fritexttermer för varje block. Några metoder för att identifiera termer är:

- Helt eller delvis använda sökstrategier som andra utformat
- I databasens tesaurus finns vanligen tips på synonyma termer. I MeSH kallas dessa "Entry terms"
- Svensk MeSH (utvecklad av, och underhålls av Karolinska Institutets bibliotek)
- En tesaurusterm kan även fungera som en fritextterm
- Analysera nyckelartiklar (guldstandard) manuellt eller med ordfrekvensverktyg, till exempel PubReMiner
- Funktionen "Related articles" i till exempel PubMed som ger tips på andra artiklar relaterade till sökresultatet
- Citeringssökning
- Fråga sakkunniga på ämnet
- Söka på internet

Avgränsningar

När man arbetar fram frågans PICO, eller motsvarande, tar man också ställning till vilka

avgränsningar som frågeställningen ska ha. Nästa fråga blir om dessa ska ingå i sökstrategin eller gallras fram vid genomgång av abstrakt.

Avgränsningar kan exempelvis vara populationens ålder, kön, språk, begränsningar i tid eller studiedesign.

Internationella databaser har inbyggda funktioner för avgränsningar, så kallade Limits. I en del databaser, som till exempel MEDLINE/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. Det gäller bland annat 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. Om Limits används i sökningen, kontrollera noga i hjälpsidorna för respektive databas vad som gäller.

Frågan om huruvida olika avgränsningar ska göras i litteratursökningen eller inte handlar till stor del om att balansera mellan att i möjligaste mån minimera risk för bias och samtidigt ta hänsyn till tidsramar och resurser. Alla beslut om avgränsningar tas i projektgruppen.

Språk

I de flesta databaser är abstrakten på engelska, även om artikeln är skriven på ett annat språk. Avgränsningar till olika språk är dock lätta att göra. En fråga som behandlats i olika studier är om man riskerar att missa studier om man inför språkbegränsningar i sökningen. Resultaten av jämförelser och utvärderingar är motstridiga. Cochranes handbok [49] hänvisar bland annat till Egger och medarbetare (1997) och Morrison och medarbetare (2012) [55] [56], och anger att språkbegränsning inte ska göras eftersom det finns en risk för att man missar studier. Flera aktuella studier visar dock på motsatt resultat när det gäller sökning på andra språk än engelska [56] [57] [58]. Det är relativt vanligt att i SBU-projekt söka med begränsning till engelska och de skandinaviska språken. Det är dock respektive projektgrupp som måste besluta om även icke-engelskspråkiga artiklar ska sökas eller om andra begränsningar ska göras.

Tidsperiod

Det kan finnas skäl att ange en begränsad tidsperiod i sökstrategin. Om sakkunniga som känner forskningsområdena väl föreslår eventuella begränsningar ska grunden till beslutet anges i projektplanen. Ett alternativ är också att alltid söka utan tidsbegränsning och göra eventuella sådana i efterhand i exempelvis EndNote, där man samlar hela sökresultatet. Begränsningar i tid handlar allt som oftast om startår. Mot slutet av projektet uppdateras sökningarna för att få ett så aktuellt sökresultat som möjligt och det är viktigt att datum för senaste sökning framgår tydligt.

Studiedesign

Projektgruppen måste också besluta om studiedesign ska ingå i själva sökstrategin eller bara vara en del av de inklusionskriterier som hanteras i abstraktgranskningen, det vill säga att man väljer att ta med eller exkludera en studie baserat på i förhand fastställda kriterier då man går igenom sökresultatets alla abstrakt. Att begränsa sökningen till studiedesign innebär, precis som med andra avgränsningar, en risk att förlora relevant litteratur. För många typer av studiedesign finns det dock utvärderade sökfilter.

Sökfilter

Sökfilter (eng. search filters, hedges) är sökstrategier som är utformade och utvärderade för att fånga en viss typ av studier, till exempel en viss studiedesign. Sökfiltren är testade mot en guldstandard av relevanta artiklar och de olika sökstrategiernas så kallade recall och precision räknas ut (se Faktaruta 4.2). Syftet är att hitta en sökstrategi som fångar så många relevanta studier som möjligt samtidigt medan antalet icke relevanta studier begränsas. Sökfiltren är utformade för att passa olika databaser men även olika versioner av samma databas kan ha olika filter. Ett filter som är gjord för PubMed passar inte Ovid MEDLINE till exempel.

Sökfiltret kombineras med sökstrategins övriga block. Eftersom nya indexeringstermer tillkommer, termer blir föråldrade med mera, behöver man regelbundet kontrollera tänkbara sökfilter. För systematiska översikter ska sökfiltren generellt sett ha hög recall det vill säga fånga så många relevanta studier som möjligt (läs mer om recall under stycke "A4.1.4.10 Att skapa sökstrategier: Litteratursökningens omfattning – en balansgång") .

En betydande samling sökfilter för olika ändamål finns vid ISSG Search Filter Resource, där informationsspecialister i InterTASC Information Specialists' Sub-Group Search Filter Resource, samlar, utvärderar och publicerar sökfilter.

Faktaruta 4.1 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. Hänsyn till resterande delar tas genom gallringen 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, det vill säga en sökning med 100 procents precision. I praktiken är detta i princip omöjligt att uppnå.

Precision och recall är två mått som man kan använda för att beskriva sökresultatet, och därför kan de räknas ut först efter att en sökning är gjord och man har granskat resultatet. Ett sökresultat kan ha mer eller mindre hög recall, och mer eller mindre hög precision, och dessa mått står nästan alltid i motsatsförhållande till varandra. När man håller på att konstruera en sökstrategi vet man naturligtvis inte hur den kan komma att prestera. Då handlar det snarare om att ha en ansats i sökarbetet som möjliggör för en viss typ av resultat, att man gör så kallade breda eller smala sökningar [59].

Faktaruta 4.2 Begrepp som används vid beskrivning av sökresultatet.

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 arti

Bred sökning

I arbetet med en systematisk översikt ska litteratursökningen vara både strukturerad och ha en uttömmande ansats. Att sökningen ska vara strukturerad innebär att den ska följa både en förutbestämd sökmetod och uppsatta kriterier (t.ex. PICO), samt genomföras i ett antal förutbestämda databaser. Till det strukturerade arbetet hör också att tillvägagångssättet är transparent och att det dokumenteras.

Att sökningen ska vara uttömmande innebär att den vid en utvärdering ska visa sig ha hög recall, det vill säga att sökningen har hittat så många som möjligt av de existerande studierna som svarar på frågeställningen. Inför en sökning vet man inte hur många relevanta studier som finns och vilka de är, vilket innebär att recall är svårt att räkna ut. Det betyder att man, när sökningen görs, inte vet hur stor andel av de relevanta studierna som faktiskt kommer att fångas. En bred sökning ökar möjligheterna att finna det mesta. Nackdelen är att ju bredare en sökning är, desto fler irrelevanta träffar kommer den att fånga, och därför blir sökningens precision lägre (se Figur 4.2). Systematiska översikter har i genomsnitt en precision på tre procent [60].

I det praktiska sökarbetet innebär detta att man i en bred sökning tar hänsyn till varierande indexering, bristande indexering, frånvaro av indexering och varierande terminologi i titel och abstrakt.

Smal sökning

För litteratursökningar som är till för andra ändamål än systematiska översikter, behöver man i sökarbetet inte sträva efter att vara lika uttömmande. Det kan exempelvis handla om sökningar till

narrativa översikter eller andra typer av kunskapssammanställningar. Det kan också handla om litteratursökningar där syftet helt enkelt är att bara hitta några bra artiklar om ett ämne och där precisionen därför väger tyngst. Vid sådana arbeten kan man alltså göra olika medvetna avgränsningar i sökningen. Det innebär att en sökning kan vara både strukturerad, det vill säga följa en noggrann metodik, och samtidigt vara precis. I boken "Systematic approaches to a successful literature review" och i en publicerad rapport beskriver Andrew Booth och medarbetare utmärkande kriterier för olika typer av översikter och utformning av litteratursökningar med olika syften [61].

Ett exempel på en väldigt smal litteratursökning, är en sökning där man söker efter två ord i artikelns titelfält och kombinerar dessa med ett booleskt "AND". En sådan sökning ger antagligen få träffar och de träffar man får bör, vid rätt val av söktermer, till stor del vara relevanta. Samtidigt missar man säkerligen stora delar av den relevanta litteraturen eftersom man inte tagit hänsyn till varierande terminologi och endast sökt efter dessa ord i titelfältet. Om de två sökorden inte är helt relevanta för frågeställningen, är det dock fullt möjligt att den smala sökningen inte alls träffar "mitt i prick" utan snarare helt utanför.

Faktaruta 4.3 Att utforma bredare sökning för systematiska översikter i jämförelse med smalare sökningar för andra ändamål.

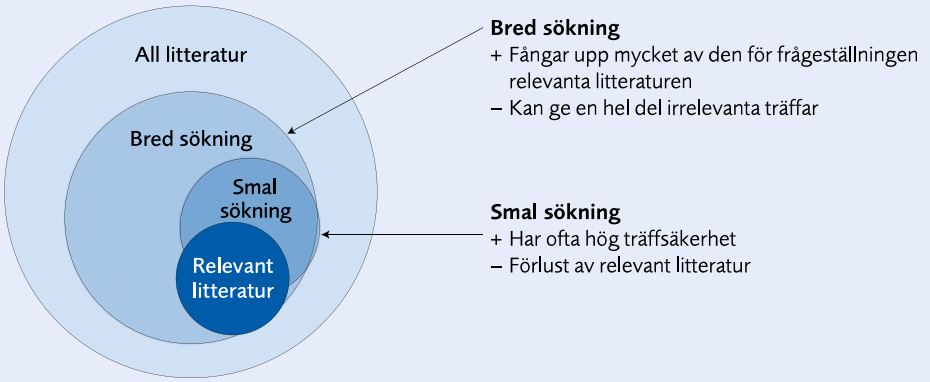
Bredare sökning för t.ex. systematiska översikter:

- 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 i PICO).
- Lägg till alternativa stavningar och böjningsformer för fritextorden.
- Trunkera fritextorden när det är tillämpligt, det vill säga sök på ordstam som slutar med ett trunkeringstecken (vanligen *). Men kontrollera och avstå om trunkeringen ger för många irrelevanta träffar.

Smalare sökning för andra ändamål än systematiska översikter:

- 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" istället för "cost*" och "qualitative study" istället för "qualitative").
- Undvik att söka på förkortningar om samma förkortning kan betyda olika saker.

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 i

beaktande antalet abstrakt man behöver granska. NNR beskriver alltså hur många abstrakt man måste läsa för att finna en relevant artikel (NNR=1/precisionen). Om projektets syfte är att besvara en frågeställning där det finns få publicerade studier, är det ganska okomplicerat att göra en bred sökning. En sådan sökning riskerar inte att missa särskilt många relevanta artiklar, samtidigt som arbetsbördan inte behöver bli så stor för dem som granskar de abstrakt sökningen identifierat.

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

Hur smal eller bred man gör sökningen är 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 jämfört 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 sökträffar med högt NNR (dvs. att 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. Men det behöver dock inte ta alltför mycket tid i anspråk att granska en abstraktlista, trots att antalet abstrakt vid första anblicken kan se ut att vara ohanterbart:

"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)" [62].

B) Litteratursökning i hälsoekonomiska utvärderingar

I SBU:s utvärderingar ingår vanligen att även utvärdera metoder ur ett ekonomiskt perspektiv. Sökstrategin för att hitta studier med ekonomiska aspekter följer i stort upplägget för att hitta det övergripande projektets studier. Ett block för populationen, ett block för interventionen och ett block med ekonomiska termer.

Ibland, till exempel vid få studier och många interventioner, kan man förenkla sökningen så att sökstrategin endast består av ett block för populationen samt ett block med ekonomiska termer.

Databaser

För ekonomiska utvärderingar inom hälso- och sjukvård rekommenderas att man gör sökningar i databaserna: Embase, HTA Database, Ovid MEDLINE/PubMed och Scopus [63][64].

För ämnesområdet socialt arbete och andra närliggande områden finns det betydligt mindre publicerat kring sökningar av ekonomiska utvärderingar, men ett exempel är ett bokkapitel av Julie Glanville och medarbetare, "Searching for evidence for cost-effectiveness decisions" [65]. För frågeställningar om insatser inom socialtjänsten gäller vanligen att sökningarna görs i samma databaser som huvudsökningen. Ibland kan man behöva lägga till en eller flera databaser med ett generellt innehåll, såsom Scopus, om den inte redan ingår i huvudsökningen. Andra tänkbara databaser med generellt innehåll som man kan behöva söka i är Web of Science eller Academic Search Elite. Man ska också kontrollera HTA-database.

Man kan även använda kompletterande metoder för att identifiera studier, såsom kontroll av referenslistor, webbsidor och olika register.

Sökfilter för hälsoekonomi

Vid hälsoekonomiska sökningar kombinerar man en ämnessökning med ett sökfilter som innehåller termer för ekonomiska aspekter. SBU använder ett filter, NHS EED, som är utvärderat och publicerat av den kanadensiska HTA-organisationen CADTH. I en utvärdering av flera filter visade sig det här filtret ha den bästa balansen mellan recall och precision [66].

Det finns ett flertal andra utvärderade relevanta hälsoekonomiska filter. Hur de presterar avseende precision och recall varierar. Om man inte behöver en uttömmande sökning finns filter med högre precision och lägre recall som man kan använda [67].

Här hittar du en lista över sökfilter för ekonomiska utvärderingar (ISSG Search Filter Resource).

C) Att söka opublicerade data och övrig grå litteratur

OBS! En arbetsgrupp på SBU ser för närvarande över hur grå litteratur, inklusive opublicerade data, ska hanteras.

Till grå litteratur räknas bland annat avhandlingar, konferenspublikationer, rapporter som inte är utgivna av kommersiella förlag.

The Third International Conference on Grey Literature in 1997* definierar grå litteratur som:

"litteratur som produceras på alla nivåer i det offentliga, på universitet, företag och industri, oavsett i vilket format, och som inte är kontrollerad av kommersiella förlag och som inte har publicering som viktigaste aktivitet"

---

* "Luxembourg Convention on Grey Literature" som antogs vid "the Third International Conference on Grey Literature 1997 " med ett tillägg 2004. http://www.greynet.org/

En särskild typ av grå litteratur är opublicerade data. Det är data från studier som antingen kan vara i form av icke-publicerade abstrakt eller fulltexter (unpublished data), eller data som finns men som man har valt att inte ta med i sina abstrakt eller fulltexter (missing data). Risken för publikationsbias har länge varit känd, det vill säga att nollresultat eller negativa resultat i kliniska studier inte publiceras i samma utsträckning som positiva resultat och att man då riskerar att föra in snedvridning av resultaten i en systematisk översikt. Det finns en risk att publicerade positiva behandlingseffekter överskattas medan negativa effekter av behandlingar underskattas om data om dessa inte publiceras i vetenskapliga tidskrifter [68][69][70][71][72][73]. Se avsnitt 6.4.5.2 för mer information om publikationsbias.

Opublicerade data kan finnas i flera olika typer av källor, till exempel konferenshandlingar, olika register för kliniska studier, läkemedelsbolagens Clinical Study Reports (CSR) och i handlingar publicerade av tillståndsgivande myndigheter.

Frågan om sökning av opublicerade data till utvärderingar av effektstudier har uppmärksammats under senare år och de flesta internationella metodböcker tar upp det som obligatoriskt eller mycket önskvärt [46][47][74][75][76]. En av anledningarna är att data som tidigare varit svåråtkomliga nu har blivit alltmer tillgängliga genom ökande krav på att pågående kliniska studier ska registreras i register och att läkemedelsbolagens tidigare svåråtkomliga Clinical Study Reports (CSR) delvis publiceras på tillståndsgivande myndigheters webbplatser. En CSR är en detaljerad beskrivning av resultaten och hur arbetet med en klinisk prövning har gått till och som lämnas av läkemedelsbolagen som underlag för bedömning av tillståndsgivande myndigheter [72][77]. Den tillståndsgivande myndigheten inom EU är European Medicines Agency (EMA) och den amerikanska motsvarigheten är US Food & Drug Administration (FDA).

Den internationella litteraturen är inte entydig om i vilken utsträckning opublicerade data kan ändra resultatet av exempelvis en metaanalys. Eftersom sökning av opublicerade data kan vara mycket resurskrävande, diskuteras också om det är möjligt att bestämma under vilka förutsättningar omfattande, detaljerade rapporter som exempelvis CSR ska sökas [69][70][73].

Konferenshandlingar och avhandlingar

I ett SBU-projekt ingår det vanligtvis inte i litteratursökningen att man söker efter konferenshandlingar. Ungefär hälften av alla studier som publiceras som ett konferensabstrakt kommer senare också att publiceras i fulltext [78]. I den vetenskapliga litteraturen finns motstridiga uppgifter om värdet av just denna publikationstyp för att undvika publikationsbias. Li och medarbetare (2017) kommer i sin genomgång fram till konferensabstrakt ofta är ofullständiga och kan innehålla motstridiga uppgifter jämfört med de publicerade artiklarna och kan därför vara vilseledande [79]. Scherer och Saldanha (2019) menar att det finns exempel på när konferensabstrakt har haft betydelse för en översikts slutresultat och inte bara för resultatets precision, och att det kan vara värt att söka efter dem särskilt när det inte finns så många studier eller om flera studier kommer fram till olika resultat [80].

Register för kliniska studier

I SBU-projekt där projektgruppen kommer fram till att det är relevant att söka i register för kliniska studier ska alltid åtminstone två register sökas: ClinicalTrials.gov och WHO:s databas ICTRP (International Clinical Trials Registry Platform). ICTRP innehåller ett antal regionala register, inklusive ClinicalTrials.gov, men har mindre utvecklade databasfunktioner. Därför ger kombinationen av dessa två register ett bättre sökresultat. Registren innehåller både pågående och avslutade kliniska prövningar. Flera studier har visat att registren bör sökas. En studie av Baudard och medarbetare (2017) visade att i 43 procent av de systematiska översikter som författarna kontrollerade kunde ytterligare RCT-studier identifieras genom sökning i register för kliniska studier. Man gjorde om 14 stycken metaanalyser, med de nya studierna inkluderade, vilket resulterade i främst en ökad precision av resultaten [81][82][83].

CSR och tillståndsgivande myndigheters data

Flera studier har kontrollerat om, och i så fall vad, opublicerade data tillför och satt resultatet i relation till resurs- och tidsåtgång. Schmucker och medarbetare (2017) kom fram till att opublicerade data har en oklar betydelse för metaanalysers resultat i medicinsk forskning och därför måste översiktsförfattare värdera om det resurskrävande arbetet med att söka opublicerade data ska göras över huvudtaget [84]. Halfpenny och medarbetare (2016) kom fram till en liknande slutsats när det gäller sökning i olika källor som register för kliniska prövningar (CSR) och i myndigheters handlingar. Eftersom det är mycket resurskrävande att söka i alla källor, rekommenderar författarna att arbetet ska ske stegvis med en noggrann genomgång av sökresultat. Rekommendationen är att börja med att söka i register, därefter handlingar från tillståndsgivande myndigheter och till sist de omfattande och detaljerade CSR-rapporterna [85]. I en annan studie av Jefferson och medarbetare (2018) identifierade författarna kriterier för när mer resurskrävande sökningar som till exempel sökningar av CSR ska göras. Bland kriterierna finns kostnaden för interventionen, sjukdomsbörda (eng. burden of disease), antal människor som kommer att kunna använda produkten, om produkten är ny, om läkemedelsgruppen är ny, eller om en stor del av RCT-studierna är finansierade av läkemedelsbolagen [72].

Sökning av grå litteratur inom socialt arbete och andra tvärvetenskapliga ämnesområden

Inom ämnesområden som socialt arbete, och överhuvudtaget inom olika tvärvetenskapliga områden, kan det finnas anledning att söka grå litteratur eftersom det inte alltid är den vetenskapligt granskade artikeln som är den självklara publikationstypen. I en utvärdering gjord på SBU framkom att identifieringen av 'genomförbarhetsstudier' kring flera av de utvärderade insatserna var av värde, även om alla inte var av hög kvalitet, liksom att projektet fick en överblick över vilka rapporter som gjorts i Sverige via den grå litteratur som identifierades. Det är också ett sätt att kartlägga var det saknas studier [86]. Erfarenheterna från SBU har också stöd i litteraturen. Adams och medarbetare (2016) [87] menar att den grå litteraturen kan ge viktig information som handlar om sammanhanget: hur, varför och för vilka en insats kan vara effektiv. En annan studie av Mahood och medarbetare (2014) [88] framhåller att sökning av grå litteratur kan ge en överblick över vilka insatser som finns för ett visst problem, vilka utvärderingar som har gjorts och inom vilka områden studier saknas. En annan viktig aspekt som tas upp i studien är att sökningen av grå litteratur kan vara utmanande när det gäller att upprätthålla den systematiska översiktens krav på att litteratursökningen ska vara systematisk, transparent och reproducerbar.

Hittills har projekt på SBU, när det varit relevant, vanligen sökt svenska avhandlingar och myndighetsrapporter. Därutöver har svenska och internationella utvärderingar av specifika namngivna insatser sökts.

Även kompletterande metoder används för att identifiera studier, framförallt kontroll av referenslistor.

D) Att söka studier med kvalitativ ansats

Hur litteratursökningarna utformas till synteser med kvalitativ ansats, är helt beroende av vilken typ av syntes (se avsnitt 3.4.3) projektet väljer att göra. Booth och medarbetare publicerade 2016 ett stöd för val av syntesmetod, och ett stöd för planering av sökning av studier med kvalitativ ansats. Stödet kallas 7S och står för Sampling, Sources, Structured questions, Search procedures, Strategies, Supplementary searching och Standards of reporting [89]. Om syntesen syftar till att beskriva ett fenomen är det av stor vikt att alla relevanta studier identifieras, då kommer litteratursökningen att vila på samma grund som den systematiska översikten. Om syftet däremot är att tolka data eller att generera teori, då kan sökningen också vara mer uttömmande, men den kan också vara upplagd på ett iterativt sätt [90].

SBU använder sig av uttömmande sökningar. Vilka sökblock som används är beroende av frågeställningen.

För frågor som handlar om erfarenheter och upplevelser av att leva med ett visst tillstånd eller om bemötande kan sökstrategin bestå av ett block för population och ett block med söktermer för erfarenheter och upplevelser, respektive bemötande. Om en smalare sökning ska göras kan ytterligare ett block med termer för studiedesign läggas till. Andra frågor kan handla om erfarenheter och upplevelser av en intervention, eller upplevelser av ett tillstånd eller en särskild intervention. Frågeställningen kan också handla om professionens attityder, erfarenheter eller upplevelser.

E) Hinder vid utformandet av sökstrategier med kvalitativ ansats inom det samhällsvetenskapliga området

Att utforma sökstrategier för att identifiera studier med kvalitativ ansats inom det samhällsvetenskapliga området (socialt arbete inkluderat) kan vara mer tidskrävande än strategier med kvantitativ ansats. Det kan bero både på författarnas sätt eller forskningsområdets tradition att namnge en studie och skriva abstrakt. Det kan saknas viktig information, sett ur ett sökperspektiv. Det kan exempelvis gälla information om studiedesign men också att begreppsbildningen kan vara heterogen. Andra orsaker kan vara brister i en databasindexeringen av kontrollerade ämnesord, eller att tesaurusen i en databas innehåller för få kontrollerade ämnesord inom området [91][92].

Databaser och andra informationskällor

Det är alltid frågeställningen som styr valet av databaser, oavsett vilken typ av studier som är i fokus. Om sökningarna av studier med kvalitativ ansats är en del av en systematisk översikt, utförs litteratursökningen i samma databaser som huvudsökningen, ofta med tillägg av databaserna CINAHL och Scopus [91]. PsycINFO kan utgöra ytterligare ett tillägg. Ibland utförs en kompletterande och samtidig sökning med fritexttermer i de olika databaser som SBU har från en databasvärd, oftast EBSCO.

Frågan om även grå litteratur ska ingå i sökningen är beroende av frågeställningen och är en fråga för projektgruppen att diskutera. Frandsen och medarbetare (2019) utvärderade nio databasers unika referenser och kom fram till att ProQuest Dissertations and Theses Global innehöll ett relativt stort antal unika referenser som sedan användes i analysen [91][93].

Här hittar du en lista över sökfilter för sökning av studier med kvalitativ ansats finns på (ISSG Search Filter Resource).

Även kompletterande metoder används för att identifiera studier, framför allt kontroll av referenslistor.

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

CINAHL
CINAHL (Cumulative Index to Nursing and Allied Health Literature) är en databas över artiklar om omvårdnad, sjukgymnastik, arbetsterapi etcetera. Den innehåller cirka 6 miljoner referenser ur cirka 5 500 tidskrifter (2019). Databasen tillhandahålls av EBSCO och åtkomst är avgiftsbelagd.

Cochrane Library
Cochrane Library består av flera olika deldatabaser. Förutom Cochrane Database of Systematic Reviews som innehåller de egna systematiska översikterna, finns även Protocols, Cochrane Central Register of Controlled Trials (Central).

Embase
Embase är den andra stora databasen inom medicinområdet. Embase innehåller cirka 32 miljoner referenser från 8 500 tidskrifter (2019). I Embase finns också innehållet i databasen MEDLINE (MeSH-databasen finns däremot inte). 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. 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.

PsycINFO
PsycINFO är en databas inom psykologi, beteendevetenskap och näraliggande ämnesområden. Databasen ger referenser till över 4 miljoner vetenskapligt granskade artiklar ur cirka 2 500 tidskrifter (2019), böcker och dissertationer. PsycINFO är avgiftsbelagd och produceras av American Psychological Association (APA). PsycINFO tillhandahålls av olika leverantörer.

PubMed
PubMed innehåller cirka 30 miljoner referenser till artiklar och ett urval fulltextartiklar från mer än 5 000 biomedicinska tidskrifter (2019). 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 tesaurus MeSH (Medical Subject Heading). Förutom dessa finns ett växande antal artiklar i PubMed som väntar på indexering. Databasen produceras av National Library of Medicine i USA och är kostnadsfritt tillgänglig via internet.

SocINDEX

SocIndex innehåller 2,5 miljoner referenser (2019), och täcker alla sociologins delar såsom antropologi, kriminologi, socialpsykologi, socialt arbete, missbruk och välfärd. Databasen innehåller referenser till tidskriftsartiklar, böcker och konferenshandlingar. Databasvärd är EBSCO och den är avgiftsbelagd.

Sociological Abstracts

Sociological Abstracts indexerar internationell litteratur inom sociologi och näraliggande ämnesområden. Databasen innehåller referenser till tidskriftsartiklar, böcker, konferenshandlingar och avhandlingar. Databasvärd är ProQuest och den är avgiftsbelagd.

Social Services Abstracts

Social Services Abstracts vetenskaplig litteratur inom socialt arbete och välfärd. Databasvärd är ProQuest och den är avgiftsbelagd.

4.1.4 Huvudsökning

När sökstrategin är genomarbetad utförs huvudsökningar. De flesta internationella metodböcker, till exempel ”Cochrane Handbook for Systematic Reviews of Interventions” [47] och ”Developing NICE guidelines” [94], anger att det inte är tillräckligt att bara söka i en databas när syftet är att hitta alla studier som besvarar frågeställningen. För att undvika risk för snedvridning av översiktens resultat genom att artiklar missas, måste flera databaser sökas, något som också stöds av studier [95] [96] [97]. Interventioner inom socialt arbete och det beteendevetenskapliga området är ofta multidisciplinära. Då kan det krävas mer specifik kunskap om både olika databaser och vilka val av databaser och andra källor som kan vara lämpliga för projektet, vilka, och hur många databaser, som är lämpliga att söka i projektet beror helt på frågeställningens ämne.

Det kan vara en fördel att börja söka i den databas som har den mest detaljerade ämnesordslistan för frågeställningen, eftersom det är den sökningen som stäms av med de sakkunniga.

Även om en artikel finns i en databas innebär det inte att den är lätt att få fram med den utvecklade sökstrategin. Eftersom samma referens kan vara indexerad (”taggad”) på olika sätt i olika databaser kan kompletterande sökningar vara värdefulla även av den anledningen. Men, att söka i flera databaser kan inte kompensera för bristfälliga sökstrategier.

Det går inte att använda samma sökstrategi rakt av för olika databaser, eftersom olika databaser har olika krav på format för sökstrategin. Därför är nästa steg att anpassa den till de övriga databaserna man beslutat om, enligt projektplanen. Sökstrategier inom hälsoekonomi och etikområdet formuleras också och sökningar utförs i de projekt där sådana aspekter ingår. Här söks i första hand de databaser som redan fastställts i projektplanen, men kompletterande databaser kan behövas.

4.1.5 Val av databaser och kompletterande söksätt

4.1.5.1 Val av databaser

På SBU söker informationsspecialisterna vanligen i minst tre databaser. För frågor inom hälso- och sjukvårdsområdet kan det räcka med sökningar i Ovid MEDLINE eller PubMed, Embase och i Cochrane Library. För multidisciplinära frågor och för frågor inom socialt arbete används vanligen PsycINFO, SocINDEX och/eller Sociological Abstract/Social Services Abstracts, samt Ovid MEDLINE eller PubMed. Scopus har ofta visat sig vara en bra kompletterande databas, och andra kompletterande databaser kan också tillkomma beroende på frågeställning.

Se Faktaruta 4.4 för exempel på databaser som är viktiga för systematiska

översikter inom hälso- och sjukvårdsområdet, respektive socialt arbete.

4.1.5.2 Kompletterande söksätt

Även om relevanta artiklar huvudsakligen identifieras i elektroniska databaser, så behövs också kompletterande söksätt. Den viktigaste metoden är ofta att gå igenom referenslistorna i relevanta systematiska översikter och primärstudier [98] [99]. Sökningarna kan vid behov även kompletteras med citeringssökningar i Scopus och den kostnadsfria Google Scholar. En annan stor licensierad citeringsdatabas är Web of Science. Cooper och medarbetare har gjort en genomgång av för- och nackdelar med olika kompletterande sätt att hitta relevanta studier [100].

Det är viktigt att redovisa vilka kompletterande söksätt som har använts.

4.1.6 Uppdateringssökning

Ofta behövs en uppdateringssökning i slutet av projektprocessen för att säkerställa att inga nya relevanta studier tillkommit under projektets gång. Riktmärket för SBU är att de slutgiltiga sökningarna ska vara högst 12 månader gamla, helst utförda inom 6 månader före publicering.

4.1.7 Sökdokumentation

SBU dokumenterar alla sökningar för respektive databas. Dessa är tillgängliga på SBU:s webbplats www.sbu.se i anslutning till rapporten. Sökarbetet beskrivs i rapporternas metodavsnitt.

Läs mer om sökdokumentation nedan.

Sökdokumentation och beskrivning av litteratursökningen som del av SBU-rapportens metodavsnitt

Två bärande principer i arbetet med systematiska översikter är att transparens och reproducerbarhet ska genomsyra hela arbetsprocessen. Det betyder att sökdokumentation och annan information om hur arbetet med litteratursökningen har utförts ska finnas tillgängligt för den som vill ta del av den systematiska översikten. Brister i rapporteringen av arbetet med litteratursökningen har uppmärksammats inom flera områden [101][102][103]. Ett krav i PRISMA statement är att en reproducerbar sökdokumentation för åtminstone en databas ska finnas tillgänglig, tillsammans med information om vilka databaser som har använts, och att man noterat eventuella begränsningar i sökningen och sökdatum. Atkinson och medarbetare (2015) har arbetat fram en detaljerad checklista på hur arbetet med litteratursökningar kan presenteras och dokumenteras [50], se Faktaruta 4.6

Faktaruta 4.5 Uppgifter som ska finnas med för att säkerställa att sökningen är reproducerbar [50].

- Databasens namn
- Databasleverantörens namn
- Datum när sökningen gjordes
- Exakta söktermer och vilken typ av term det är, det vill säga indexeringsord eller fritext och vilka fält
- Eventuella begränsningar
- Hur termerna kombinerats med de booleska operatorerna och eventuella närhetsoperatorer utskrivna

Tabell 4.1 Exempel på SBU:s sökdokumentationsmall

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

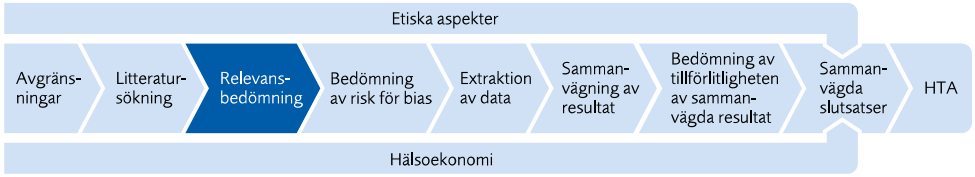
The search result, usually found at the end of the documentation, forms the list of abstracts. **[MeSH]** = Term from the Medline controlled vocabulary, including terms found below this term in the MeSH hierarchy; **[MeSH:NoExp]** = Does not include terms found below this term in the MeSH hierarchy; **[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

På SBU dokumenteras alla sökstrategier. Det gäller såväl för huvudsökningens respektive databas som för sökdokumentationen för ekonomiska utvärderingar, i förekommande fall för etiska aspekter och sökning av grå litteratur. Sökdokumentationerna publiceras som bilagor till varje rapport på SBU:s webbplats, sbu.se. Arbetet med litteratursökningen beskrivs i respektive rapports metodavsnitt: Metod för den systematiska översikten.

4.1.8 Verktyg för referenshantering

Sökresultaten importeras till ett referenshanteringsprogram där dubblettkontroll görs. I SBU:s projekt används EndNote. När alla sökningar är gjorda och alla dubbletter är borttagna återstår den manuella granskningen av de framsökta abstraktens relevans (se Kapitel 5).

5. Bedömning av relevans

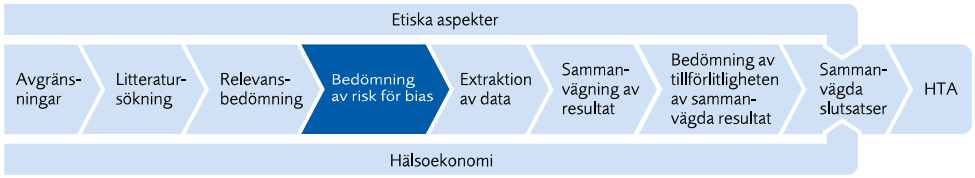


Urvalet av studier görs i flera steg och bygger på att två personer först oberoende av varandra bedömer studierna varefter de gör en gemensam slutbedömning av om en studie ska ingå eller exkluderas (konsensusförfarande).

De studier som ska ingå i frågans vetenskapliga underlag måste vara relevanta, det vill säga uppfylla PICO (eller motsvarande). Eftersom SBU vanligen tillämpar breda sökkriterier kommer sökresultaten innehålla en stor mängd referenser som inte är relevanta. I ett första steg gallras sådana studier bort redan på grundval av information från titel och abstrakt. Gallring av artikelabstrakts görs ofta först av SBU:s kansli, antingen direkt i referenshanteringssystemet EndNote eller i verktygen Rayyan eller Covidence. Här sorteras uppenbart irrelevanta studier bort. Sakkunniga får därefter ta ställning till en kortare lista och gallra vidare utifrån denna. Artiklar som av titel och abstrakt att döma skulle kunna uppfylla urvalskriterierna beställs i fulltext. Observera att man inte i sammanställer orsakerna till att abstrakts exkluderas utan det räcker att enbart notera antalet.

Vid närmare påseende kommer många av studierna inte att vara relevanta. Det kan finnas flera skäl, förutom att de inte uppfyller PICO (eller motsvarande). Studien kan visa sig ha fel publikationsformat, till exempel brev till redaktören, eller vara av diskuterande natur utan egna resultat. Dubbelpublikationer kan förekomma, det vill säga att samma studie publiceras i två tidskrifter, och då ska den ena exkluderas. När de sakkunniga kommit överens om vilka artiklar som ska exkluderas ska kansliet upprätta en förteckning över dem, samt ange orsakerna till att de har gallrats bort. Observera att endast en orsak ska anges per artikel, även om det kan finnas flera skäl till att exkludera artikeln.

6. Bedömning av risk för bias



Nästa steg i processen är att bedöma risken för att resultaten i de inkluderade studierna har snedvridits, det vill säga risken för bias. Bias kan uppstå i såväl design av studien, som i dess genomförande. Det kan också vara svårt att avgöra hur säkra resultaten är om studien har rapporterats på ett bristfälligt sätt. Det finns många sorters bias, där en del förekommer oavsett studiedesign, medan andra är viktiga att undersöka för vissa typer av studiedesign. För den som är intresserad av att veta mera om bias av olika slag finns en webbsida som beskriver ca 60 olika typer av bias.

Idag finns det internationella riktlinjer för hur studier ska genomföras och rapporteras, till exempel Consolidated Standards of Reporting Trials (CONSORT) [104] för randomiserade studier, Standards for Reporting of Diagnostic Accuracy (STARD2015), [105] för studier om diagnostisk tillförlitlighet och Consolidated criteria for REporting Qualitative research (COREQ) [106] för studier med kvalitativ metodik. De senaste versionerna av riktlinjerna finns samlade på hemsidan för det internationella initiativet Enhancing the QUAlity and Transparency Of health Research (EQUATOR).

Tidigare bedömdes studiens *kvalitet*. Nackdelen är att begreppet studiekvalitet inte tar hänsyn till att olika utfall kan vara olika känsliga för brister i design och genomförande. En studie som till exempel rapporterar såväl mortalitet som skattning av symtom kan ha olika risk för bias i de olika utfallen. Studiekvalitet har därför ersatts med risken för att ett resultat snedvridits (eng. risk of bias). För studier med kvalitativ metodik är terminologin inte lika utvecklad och SBU använder begreppet ”risk för att metodbrister påverkar fynden”.

Bedömning av risk för bias i utfallen innehåller med nödvändighet subjektiva inslag och därför är det viktigt att vidta mått och steg för att minska risken för subjektivitet. För det första ska minst två personer granska studierna, först oberoende av varandra och därefter görs en samordnad bedömning. För det andra ska standardiserade checklistor, mallar, användas som stöd för bedömningen. Den självständiga bedömningen görs oftast i fyra steg (Figur 6.1). SBU har valt de checklistor som används inom Cochrane Collaboration och översatt dem till svenska. Undantagen är checklistorna för studier med kvalitativ metodik som utvecklats av SBU eftersom det för närvarande saknas en internationellt vedertagen mall som bygger på begreppet risk [107].

Figur 6.1 Den individuella bedömningen av en studie, med hjälp av granskningsmall, sker i fyra steg. Resultaten av denna bedömning stäms sedan av med en eller flera andra granskare i projektgruppen.

Samtliga mallar är uppbyggda på ett likartat sätt. De består av ett litet antal riskområden (eng. domains) som vart och ett representerar en typ av risk för bias, till exempel selektionsbias och bias som följd av selektiv rapportering. För varje riskområde finns hjälp i form av ett antal stödfrågor (eng. signalling questions). Att fastställa risken för bias görs såväl utifrån svaren på stödfrågorna, som vid en bedömning av vad eventuella brister betyder för utfallet, eftersom vikten av en risk kan variera både beroende på forskningsområde och kontext. Den erfarne granskaren kan välja att inte fylla i de enskilda svaren på stödfrågorna utan bedöma risken per riskområde direkt.

Klassificeringen av risk varierar mellan mallarna och graderas vanligen mellan hög och låg risk. För icke-randomiserade studier finns även nivån oacceptabelt hög risk. Studier som man bedömer ha oacceptabelt hög risk för bias inom ett riskområde tas omedelbart bort från vidare analys.

Granskningen av risk för bias avser dels risken per riskområde, dels den övergripande risken för bias. Huvudprincipen för SBU är att resultat med totalt sett hög risk för bias inte ska ingå i det vetenskapliga underlaget. Studier där resultaten har hög risk för bias ska sammanställas i en separat lista. I vissa fall kan dock projektgruppen besluta att även ta med resultat med hög risk. Detta ska då framgå redan i projektplanen och motiveras. Ett skäl till att inkludera studier med hög risk är exempelvis om forskningsfältet är nytt och att antalet studier sannolikt är få.

Denna del av metodboken beskriver översiktligt dels vilka risker som finns för originalstudier av olika typer som är vanliga i SBU-rapporter, och dels hur mallarna ska användas. Texten ska alltså ses om en bakgrund och ett komplement till de detaljerade instruktioner som finns för respektive mall.

6.1 Risk för bias i interventionsstudier (RCT och NRSI)

SBU-projekt som utvärderar effekter av interventioner bygger oftast på studier som har en eller flera kontrollgrupper, med eller utan slumpmässig (randomiserad) gruppindelning av studiedeltagarna. Detta avsnitt beskriver hur man systematiskt bedömer risken för bias i både randomiserade studier (RCT, randomised controlled trials) och i icke-randomiserade studier, där deltagarna alltså inte har fördelats slumpmässigt mellan grupperna (NRSI, non-randomised studies of interventions), se Tabell 6.1.

SBU har valt att modifiera och översätta granskningsmallar för RCT- respektive NRSI-studier, som har utvecklats av universitetet i Bristol i samarbete med Cochrane Collaboration.

SBU:s granskningsmall för *randomiserade studier* bygger på ROB 2 (Risk of Bias tool 2) [108]. ROB 2 finns i flera versioner, för parallella grupper med individuell randomisering, respektive randomisering på gruppnivå (t.ex. hela skolor, s.k. klusterrandomisering) samt för överkorsningsstudier. ROB 2-mallarna och mycket detaljerade instruktioner för hur de ska användas finns på universitetet i Bristols webbplats. SBU har enbart översatt mallen för individuell randomisering och parallella grupper.

Granskningsmallen för *icke-randomiserade studier* bygger på ROBINS-I (Risk of bias in non-randomised studies of interventions) [109]. ROBINS-I och anvisningar för hur den ska användas finns här. Granskningsmallen kan användas för såväl prospektiva som retrospektiva studier, inklusive registerstudier.

Båda mallarna finns i en version för ITT-analys (att tilldelas en intervention) och en version för per protokoll-analys (att fullfölja en intervention). I SBU-rapporter användas oftast versionen för ITT-analys. Enda skillnaden mellan de två versionerna är bedömning av avvikelser från planerad intervention.

6.1.1 Steg 1: Bedömning av riskområden

De aspekter som tas upp i det första riskområdet (dvs. domän 1), om fördelning mellan grupperna, är den stora skillnaden mellan en RCT och en NRSI. De övriga riskområdena är gemensamma för båda studietyperna, se Tabell 6.1. Riskområde 6 om jäv och intressekonflikter finns inte med i de ursprungliga granskningsmallarna (ROB 2 och ROBINS-I) utan är ett tillägg av SBU.

Tabell 6.1 Riskområde 1–6 i granskningsmallarna för bedömning av risk för bias för RCT-, respektive NRSI-studier.

Riskområde	RCT	NRSI
1.	Gruppindelning: Randomisering	Gruppindelning: A) Confounders B) Selektion C) Klassificering av deltagare och interventioner
2.	Avvikelser från planerade interventioner	
3.	Bortfall	
4.	Mätning av utfallet	
5.	Rapportering	
6.	Jäv och intressekonflikter	

6.1.1.1 Riskområde 1: Bias som en följd av gruppindelning (RCT)

I studier där man jämför en eller flera interventions- och kontrollgrupper med

varandra bör deltagaregenskaper, till exempel ålder och kön, vara jämnt fördelade mellan grupperna.

Styrkan med randomisering är att den förebygger den bias som uppstår i samband med att deltagarna delas in i grupper. En välgjord randomisering gör att både de deltagaregenskaper man känner till och de man inte känner till fördelas slumpmässigt mellan interventions- och kontrollgrupperna. Om antalet deltagare är tillräckligt stort blir det en jämn spridning av deltagaregenskaper i grupperna. Okända prognostiska faktorer som kan förutsäga utfallet (t.ex. svårighetsgrad av sjukdom eller samsjuklighet) blir balanserade. Bedömningen av risk för bias grundar sig därför på hur forskarna genererat sekvensen för slumpmässig gruppindelning och om sekvensen har kunnat påverkas.

Det finns flera sätt att ta fram en slumpmässig sekvens som styr vilken gruppdeltagarna ska tilldelas, alltifrån enkla manuella metoder som att singla slant eller använda slumpgeneratorer till webbaserade program för randomisering. Det viktiga är att det inte ska finnas någon möjlighet att påverka sekvensen. Ibland läggs begränsningar in i processen för att få jämna proportioner mellan grupperna (t.ex. 1:1). Sådana begränsningar kan påverka effekten av randomiseringen.

Nedan kan du läsa mer om olika typer av randomisering.

Olika typer av randomisering: blockrandomisering, stratifiering och minimisering

Blockrandomisering

Vid framför allt små studier finns det en risk för att randomiseringen leder till att grupperna blir olika stora, något som medför att den statistiska styrkan minskar. Med blockrandomisering kan man undvika detta och möjliggöra för grupperna att bli lika stora. Blockstorleken kan vara fastställd i förväg, till exempel fyra eller åtta personer, eller vara slumpmässig. Inom varje block kommer lika många personer att fördelas till varje grupp. Om blockstorleken exempelvis är fyra och studien har två grupper kommer två personer att fördelas till vardera gruppen. Nackdelen med en fördefinierad blockstorlek är att det finns en risk för att grupptillhörigheten blir förutsägbar.

Stratifiering

Om man vill säkra god jämförbarhet mellan grupperna kan man vid urvalet använda sig av stratifiering. Deltagarna indelas i strata efter vissa prognospåverkande faktorer. Exempelvis kan personer över 50 år utgöra ett stratum och de som är 50 år eller yngre ett annat stratum. Randomisering sker separat inom varje stratum. Om en undersökning har tillräckligt stort antal deltagare kan man göra en jämförande analys av resultaten dels i varje grupp, dels i varje stratum inom varje grupp. Vilka jämförelser som ska göras ska specificeras i förväg i studiens protokoll.

Minimisering

Tanken med minimisering är att den första studiedeltagaren slumpas till någon av studiens grupper. Därefter hamnar var och en av de följande försökspersonerna i den grupp som (med hänsyn till alla identifierade confounders) mest jämnar ut obalansen mellan grupperna.

Om man känner till principen bakom randomiseringen kan det vara möjligt att påverka vilka individer som hamnar i vilka grupper, därför skall sekvensen vara dold för de inblandade i studien tills interventionen startar (dold allokering). Det

effektivaste sättet att dölja gruppindelningen är att tilldelningen sköts av en tredje part.

Ibland kallar man metoder som baseras på till exempel födelsedatum, veckodatum eller datum för besök hos läkare för kvasi-randomiserade. Dessa metoder är varken randomiserade eller lämpliga eftersom det går att styra gruppindelningen.

Den första mätningen av deltagarnas olika egenskaper kallas ofta baslinjemätning och resultaten från den bör presenteras i en tabell. Mätningen bör vara gjord före randomiseringen om interventionen inte kan blindas. Baslinjemätningen är en viktig källa till kunskap när man vill granska randomiseringen. Om det finns skillnader mellan grupperna kan randomiseringen ha misslyckats. Små skillnader kan bero på slumpen men man bör se upp med ovanligt stora skillnader i gruppstorlek och deltagaregenskaper, överdriven likhet mellan grupperna och att viktiga egenskaper saknas.

Du som vill fortsätta läsa om endast RCT kan hoppa fram till avsnitt 6.1.1.3 ”Avvikelser från planerade interventioner”.

6.1.1.2 Bias i samband med gruppindelning (icke-randomiserade studier) – riskområde 1A, 1B, 1C

Studier utan randomisering har fördelen att de kan genomföras även då en randomisering skulle vara praktiskt omöjlig eller oetisk att genomföra. Men deltagaregenskaperna blir av olika skäl ofta ojämnt fördelade mellan grupperna, vilket kan leda till hög risk för bias. Resultaten från en NRSI studie bör därför ställas i relation till vilka resultat som skulle ha uppnåtts om studien varit randomiserad. Cochrane Collaboration rekommenderar att man definierar en idealisk randomiserad studie, utan de eventuella hinder som kan finnas för en välgjord RCT (praktiska, etiska eller ekonomiska). Hur skulle populationen väljas? Hur skulle interventionen ges? Detta kan vara omständligt att göra för alla ingående studier men kan vara lämpligt att göra för åtminstone en, eller några, av studierna i projektet, till exempel som en gruppövning i projektgruppen.

Exempel på bias som en följd av gruppindelning

Det går att fördela studiedeltagarna så att det blir balans mellan egenskaper som är kända, såsom ålder och kön (s.k. matchning). Däremot går det inte att ta hänsyn till egenskaper hos deltagarna som är okända eller svåra att mäta. Ett exempel är de stora kohortstudier som genomfördes i slutet av 1980-talet som visade att östrogenbehandling efter menopaus (HRT, hormone replacement therapy) minskade risken för hjärt- och kärlsjukdom. När man undersökte effekten i randomiserade prövningar såg man dock inga statistiskt säkerställda skyddande effekter av östrogenbehandlingen. Detta tolkades som att de kvinnor som i kohortstudierna behandlades med östrogentillägg hade en mer hälsomedveten livsstil och därmed en bättre prognos, vilket innebär att det var selektionen av kvinnor som påverkade resultatet.

Riskområde 1A: Identifiering och kontroll av confounders

Confounders, ibland också kallade störfaktorer, är prognostiska faktorer som både påverkar vilken grupp en individ hamnar i och utfallet. Confounders är

egenskaper och karakteristika hos populationen som kan komma att bli ojämnt fördelade mellan de grupper som ska jämföras. Vanliga confounders är svårighetsgrad av sjukdom, samsjuklighet, ålder och socioekonomiska faktorer. Projektgruppen behöver komma fram till vilka confounders som kan vara väsentliga för forskningsfrågan och sammanställa dem innan man börjar granskningen. Detta görs lämpligen i samband med ett projektmöte. Det är en styrka om valet av confounders baseras på tillförlitliga forskningsdata.

Confounders som ser obetydliga ut vid studiestarten och den första mätningen av deltagaregenskaperna kan få konsekvenser senare i studien. Confounding som varierar över tid (eng. time-varying confounders) beror på faktorer som förändras efter det att interventionerrna har startat.

Det händer också att faktorer som kommer att användas i analysen kan läggas till av forskarna när baslinjemätningen redan är avslutad och interventionen har startat (eng. postintervention variables). I en prospektiv studie kan det bero på brister i planeringen eller på att nya egenskaper hos deltagarna har noterats (exempelvis reaktion på behandlingen). Effekten av interventionen ska beräknas på variabler som mätts upp vid baslinjen innan interventionen. Om nya variabler tas in efter att interventionen har startat kan de inte bedömas som om de har påverkats av den.

När confounders är kända kan man ta hänsyn till dem genom att kontrollera för dem. Förutsättningen är att man har tillgång till valida och reliabla data. När sådana saknas kan de i vissa fall ersättas av data som motsvarar confoundern, exempelvis viktnedgång, för att kontrollera för allvarlighetsgrad av tillstånd eller utbildningsnivå, och inkomst för att kontrollera för socioekonomisk status, se nedan.

Det är sällsynt att en NRSI har en låg grad av confounding. När man bedömer risken för bias i studier där forskarna har kontrollerat för confounders är det därför viktigt att bedöma om kontrollen var tillräcklig och om det är risk för att det finns kvarvarande confounding (eng. residual confounding). Kvarvarande confounding kan antingen vara helt okänd eller känd men inte mätt.

Läs mer om hantering av confounders nedan.

Hantering av confounders

Lämpliga metoder för att kontrollera för *uppmätta* confounders är stratifiering, regression, matchning och invers probabilitetsviktning (se nedan). I en del fall kan man använda "negativa kontroller" för att mäta om effekten fanns i gruppen innan interventionen. Varje metod förutsätter att det inte finns några confounders som inte har blivit uppmätta eller några kvardröjande effekter av confounders (eng. residual confounding).

Residual confounding är den kvarvarande kumulativa effekten av confounders som man inte har kunnat hantera eller justera för. Om denna effekt finns och vilken betydelse som kvarvarande confounders har är en tolkningsfråga. Om man bedömer att det finns kvarvarande confounding, vill man gärna veta i vilken tänkt riktning denna leder: till överskattning eller underskattning av effekt.

Det finns metoder (t.ex. directed acyclic graphs, DAGs) för att reducera antalet kontrollerade confounders eftersom det annars leder till en överjustering, alltså en introduktion av bias. Det att man justerat för många confounders i en exempelvis en regressionsanalys, behöver inte nödvändigtvis alltid vara en bra sak.

Riskområde 1B: Selektionsbias

Selektionsbias inträffar när några deltagare, eller den första uppföljningen för några deltagare eller några utfall exkluderas på ett sätt som leder till att sambandet mellan intervention och utfall störs. Det innebär att selektionen är relaterad till både intervention och utfall. Till skillnad från confounders uppstår selektionsbias som en följd av brister i studieprocessen. Det finns flera typer av selektionsbias och de kan uppstå både i baslinjen och efter det att interventionen påbörjats.

Olika typer av selektionsbias

- **Lead time bias:** Uppstår när upptäckten av ett tillstånd räknas som startpunkt för sjukdomsförloppet. Vissa tillstånd behöver lång tid på sig för att utvecklas till märkbara symtom. Om man till exempel mäter livslängd från upptäckt med screening innan symtom har uppstått kan livslängden därför bli längre än om man mäter tiden från symtomdebut.

- **Immortal time bias:** Uppstår när analysen även innehåller utfall som inte kan vara ett resultat av insatsen eftersom de har uppmätts under fel tidsperiod. Ett exempel är om utfallet inträffar före insatsen och rapporteras som resultat: Som utfallet "död" i en studie där tiden mellan gruppindelning och insats är så lång att de sjukaste deltagarna hinner dö innan man påbörjat insatsen. Om dessa räknas in som "icke behandlade" leder det till en selektiv felklassificering. Ett annat tillfälle då immortal time bias kan uppstå är i studier med långa uppföljningstider:
  - **Exempel:** I en studie vill man studera hur stor andel utskrivna patienter som återinsjuknar efter en viss behandling. Resultatet mäts med läkemedelskonsumtion. Utfallet är "antal patienter som tar ut preparat X från apoteket inom 50 dagar efter avslutad behandling". En patient som tar ut preparatet dag 45 kategoriseras som exponerad för preparatet, en patient som dör innan hen tar ut det kategoriseras om icke exponerad för preparatet. Detta leder till en selektiv felklassificering in i gruppen "icke exponerad". Immortal time bias kan identifieras genom att jämföra studien med den ideala RCT:n där patienterna skulle följas från randomisering även om insatsen inträffade något senare.

Ett exempel är en studie av balansträning för att förebygga fallolyckor bland äldre. Om man väljer bort de äldsta deltagarna i interventonsgruppen riskerar man stora selektionsbias. Det beror på att man förlorar data för de deltagare som både har den största risken att råka ut för en fallolycka och minsta möjligheten att genomföra ett träningsprogram. Att välja bort de äldsta kan dels göras i baslinjen för att man ser det som alltför ansträngande för dem att delta eller senare för att man exempelvis upptäckt fler skador hos dem än hos övriga

deltagare.

Det finns statistiska metoder, till exempel invers probabilitetsviktning, som kan skydda mot selektionsbias men ofta saknas data för att genomföra sådana. Därför klassificeras vanligen risken för bias som hög eller oacceptabelt hög.

Mer om invers probabilitetsviktning

Invers probabilitetsviktning (eng. inverse probability weighting, IPW) är en statistisk teknik där resultat från grupper som är underrepresenterade i en studie (jämfört med hur vanligt förekommande dessa grupper är i den population man är intresserad av att undersöka) ges en ökad vikt i analyserna. Den ökade vikten för de underrepresenterade gruppernas resultat ska motsvara deras egentliga andel av den population som man undersöker

Den här tekniken kan användas dels om det saknas möjligheter att testa den egentliga populationen vilken man är intresserad av, dels om delar av den population man är intresserad av faller ur analysen för att det saknas data.

Riskområde 1C: Klassificering av deltagaregenskaper och interventioner

Denna typ av bias uppstår främst i retrospektiva studier. I prospektiva kontrollstudier har man ofta en liknande ordningsföljd för arbetsprocessen som i randomiserade studier:

rekrytering → gruppindelning → baslinjemätning → interventionen → uppföljning och datainsamling → analys

och skillnaden ligger då främst i metoden för gruppindelning, medan ordningsföljden i en retrospektiv studie däremot kan avvika betydligt:

datainsamling av deltagaregenskaper och resultat från register eller journal → gruppindelning baserad på deltagaregenskaper såsom diagnos, exponering eller vilken intervention man får → analys.

Det betyder att man i retrospektiva studier kan ha tillgång till både deltagaregenskaper och resultat innan gruppindelningen börjar. Om man dessutom ska ta in data från flera typer av register eller journaler kan fel uppstå i sammanställningen av dem.

I retrospektiva studier är det därför viktigt att beskrivningen av grupperna är så tydlig att risken för felklassificering blir minimal. Viktiga uppgifter kan vara interventioner typ av intervention, dos, dosering eller antal sessioner, behandlingstid och tidpunkt för interventionen.

Felklassificering behöver inte leda till bias. Bias uppstår om felklassificeringen påverkas av utfallet (eng. differential misclassification). Risken för bias minskar om data som används till klassificeringen samlas in innan resultatet är känt. Om det inte är möjligt kan man i vissa fall samla data på ett sätt som förhindrar att man får kännedom om interventionen och resultat innan klassificering och gruppindelning, till exempel genom att data om utfallet är dolda för den som ska klassificera grupperna.

Olika typer av bias som kan uppstå vid klassificering av deltagare och interventioner

- **Information bias:** Data kommer från otillförlitliga källor och register.
- **Measurement bias:** Mätfel från felkalibrerade instrument (även frågeformulär) eller självrapporterade data.
- **Observer bias:** När den som genomför studien känner till grupptillhörigheten kan förväntanseffekter påverka tolkningen av data.
- **Recall bias:** Minnet hos dem som deltar kan vara påverkat av kännedom om nuvarande tillstånd. Ett exempel är en studie på smärtpatienter där man vill veta hur tidigare trauman påverkar utfallet av behandling. Om man frågar undersökningspersonerna om tidigare trauman för att dela in dem i grupper finns det en risk för minnet blir påverkat av situationen. Uppgifter om tidigare händelse bör vara antecknade när de skedde exempelvis i journaler.

6.1.1.3 Riskområde 2: Avvikelser från planerade interventioner (RCT och NRSI)

Om det uppstår avvikelser från de interventioner forskarna planerat att undersöka behöver man ta reda på varför de uppstått, hur stora de är och om de är jämnt fördelade mellan grupperna.

Risken för bias ökar om deltagare och/eller behandlare (forskare och personal där interventionen genomförs) känner till vilken grupp deltagaren tillhör. Kännedom om grupptillhörighet kan leda till att någon av grupperna reagerar negativt eller positivt. Ett exempel är om deltagare i kontrollgruppen förändrar sina hälsobeteenden för att kompensera för att de inte får en behandling. De som ger behandlingen kan också börja behandla deltagare olika på grund av de känner till deltagarnas grupptillhörighet, exempelvis genom att ge extra uppmärksamhet och omsorg till någon av grupperna.

Avvikelser kan också uppstå av kliniska skäl som skulle ha inträffat oberoende av om det pågick en studie eller inte. Då är den avvikelsen oftast jämnt fördelad mellan grupperna.

En annan avvikelse som kan uppstå är om deltagare byter interventionen (och deras grupptillhörighet förändras). Det kan antingen vara i linje med hur vården brukar ges till den aktuella patientgruppen, eller så kan det bero på att grupptillhörigheten inte är dold. Om en stor andel av deltagarna byter grupp kan det i randomiserade studier medföra att effekten av själva randomiseringen går förlorad. En större andel än 5 procent räknas ofta som en stor andel i det här sammanhanget.

Många tillstånd behandlas med flera behandlingar samtidigt. Efter ett benbrott kan man exempelvis bli ordinerad fysioterapi tillsammans med smärtlindrande läkemedelsbehandling. I studier kallas detta ofta för co-intervention och då syftar man på den behandling som deltagarna får samtidigt som interventionen. Risken för att co-interventioner fördelas ojämnt mellan grupperna ökar om behandlarna känner till deltagarnas grupptillhörighet.

6.1.1.4 Riskområde 3: Bortfall (RCT och NRSI)

Bortfall kan avse antingen individer eller enstaka mätpunkter. Data kan saknas av flera olika skäl, till exempel att:

- Deltagarna avbryter medverkan eller inte kan lokaliseras (eng. lost to follow-up).
- Deltagarna deltar inte i en uppföljningsmätning.
- Mätresultat förloras eller är inte tillgängliga av andra anledningar (eng. missing data).

Bias kan uppstå beroende på att bortfallet är obalanserat mellan grupperna, att orsakerna till bortfall är obalanserat, om det finns skillnad i utfall mellan dem som föll bort jämfört med dem som var kvar i studien eller vad prövarna gjort för att hantera bortfall i sina analyser.

Det finns inga regler för vad som kan anses vara ett högt bortfall. För kontinuerliga utfallsmått är det osannolikt att resultatet snedvridits om bortfallet understiger 5 procent. För dikotoma utfallsmått är risken för bias förknippad med risken för utfallet, ett lågt bortfall kan alltså leda till bias om utfallet är sällsynt.

Även om bortfallet är balanserat vad gäller storleksordningen kan orsakerna till bortfall skilja sig åt mellan grupperna, vilket också kan introducera bias. Ett exempel är om deltagare med sämre kliniska utfall har större benägenhet att avbryta sin medverkan i studien beroende på biverkningar och att detta i högre grad inträffar i gruppen som får interventionen. Då kommer utfallet att bli snedvridet till interventionens favör.

Trots bortfall kan utfallet vara robust om bortfallet hanteras på ett bra sätt i analysen och om författarna gjorde några sensitivitetsanalyser (känslighetsanalyser).

Det finns tre vanliga sätt att hantera bortfallet statistiskt:

- Ofullständiga observationer tas bort (eng. complete case analysis). Metoden riskerar dock att introducera bias.
- Imputering där saknade värden läggs in före analys. Imputerade data betraktas som bortfall.
- Analys av ofullständiga data genom en metod som inte kräver ett komplett dataset.

Mer om imputering

Metoderna Last observation carried forward (LOCF) eller Baseline observation carried forward (BOCF) som används vid imputering kan ge problem om det finns en underliggande trend till försämring eller förbättring och analyser med dessa metoder bör granskas noga.

Analys av ofullständiga data genom en metod som inte kräver ett komplett dataset

I dessa fall kan man till exempel använda sig av någon metod som baseras på Maximum likelihood (t.ex. Expectation-Maximization). Då beräknar man parametrar till sin statistiska modell baserat på de data som finns tillgängliga, och därefter räknar man fram nya värden för de datapunkter som saknas med hjälp av den nyss framtagna modellen. Detta baseras på ett antagande om multivariat normalfördelning av data.

Ett annat sätt är att använda sig av multipel inputering (eng. multiple imputation). Då beräknas ett set av möjliga värden för varje saknat värde, så att man får flera möjliga kompletta dataset. Varje dataset används sedan för att ta fram ett resultat. I sista steget kombineras alla de framtagna resultaten till ett enda gemensamt, genomsnittligt resultat, som då blir slutresultatet.

6.1.1.5 Riskområde 4: Mätning av utfallet (RCT och NRSI)

Om de som mäter utfallet är medvetna om vilken av grupperna deltagarna tillhör finns det risk för bias. Därför är det viktigt att de är blindade. Kännedom om grupptillhörigheten kan leda till att effekterna överskattas. Överskattningen blir ofta större när utfallsmåtten grundas på en subjektiv bedömning.

Den som mäter utfallet kan vara:

- *Deltagaren* när utfallet är ett så kallat deltagarrapporterat utfall, till exempel livskvalitet och poäng på en skattningsskala. Data erhålls exempelvis genom intervjuer, frågeformulär eller dagböcker. Deltagaren betraktas som bedömare även om en blindad intervjuare ställer frågor och fyller i ett formulär. Bedömningen påverkas vanligen av kännedom om interventionen.
- *Behandlaren* när utfallet är resultatet av en klinisk undersökning eller ett beslut grundat på undersökningen. Utfallet är ett beslut som görs av behandlaren. Beslutet kan vara högst beroende av kunskap om grupptillhörighet. Exempel är sjukhusinläggning, avsluta behandling och remittering. Bedömningen påverkas vanligen av kännedom om interventionen.
- En *observatör* som inte är direkt inblandad i interventionen. Om utfallet inte innebär någon bedömning påverkas det vanligen inte av kännedom om insatsen. Exempel på utfall är mortalitet oavsett orsak. Om utfallet kräver en viss grad av bedömning, till exempel granskning av röntgenbilder och kliniska händelser förutom död, påverkas utfallet vanligen av kännedom om interventionen.

När den som mäter utfallet inte är blindad och utfallet kan påverkas av kännedom om grupptillhörighet finns flera saker att ta hänsyn till vid bedömning av risk för bias. Exempel på sådant man bör beakta är grad av förväntningar eller preferenser hos den som mäter utfallet, grad av medverkan i deltagarens vård och påverkan från andra parter i studien.

6.1.1.6 Riskområde 5: Analys och rapportering (RCT och NRSI)

Selektiv rapportering kan innebära att vissa utfallsmått inte rapporteras även om de har mätts. Sådana avvikelser bör man behandla i samband med att man gör en

bedömning av publikationsbias i GRADE (se avsnitt 9.6). Selektiv rapportering kan också innebära att endast utvalda mätningar eller analyser av ett utfallsmått redovisas och ingår i bedömningen av risk för bias.

För att kunna bedöma om författarna medvetet valt ut vissa mätningar eller analyser behöver man läsa studiens projektplan (protokoll) eller en statistisk analysplan. Det är viktigt att kontrollera att de publicerade analyserna verkligen var planerade. Om till exempel datum för ett dokument i ett prövningsregister ligger endast några månader innan slutlig publikation är det osannolikt att analyserna verkligen var specificerade i förväg. Om det inte går att få tag på protokollet går det fortfarande att uppskatta risken för bias genom att till exempel jämföra texten i metodavsnittet med resultatredovisningen eller om det finns flera publikationer om samma studie.

Tips på frågor som kan vara till hjälp för att bedöma risken för selektiv rapportering (randomiserade och icke-randomiserade studier) [47]

- Definierades subgrupper på ett ovanligt sätt?
- Finns det skillnader mellan publikationer i vilka deltagare som analyserades?
- Användes olika tröskelvärden för att skapa kategorier i olika publikationer av samma studie?
- Användes ett ovanligt kompositmått?
- Har subskalor aggregerats på ett ovanligt sätt?
- Finns det en skillnad mellan olika publikationer i vad som är primära eller sekundära utfallsmått?
- Gjordes en eller flera justerade analyser men ingen rapporterades?
- Gjordes analyser med imputering och redovisades de utan motivering?
- Användes flera imputeringsmetoder men bara resultatet av en redovisades?

När man överväger risk för bias ska man ta i beaktande både storlek, riktning och statistisk signifikans för estimaten. Om det finns bevis för att några mått eller analyser i en placebokontrollerad studie inte har rapporterats, men det redovisade resultatet är icke-signifikant, eller visar på nära ingen effekt, är det mindre sannolikt att studieförfattarna har valt det rapporterade estimatet baserat på dess resultat.

6.1.1.7 Riskområde 6: Jäv

Bedömningen av risken för att resultatet påverkats av intressekonflikter görs först i steg 3, som beskrivs nedan.

6.1.2 Steg 2: Sammanvägd risk för bias i ett enskilt utfall

Bedömningen avslutas med en sammanvägning av risken för bias. Den grundar sig på överväganden om hur riskerna påverkar utfallet totalt sett. Som tumregel gäller följande för låg respektive hög risk för bias:

- För att utfallet ska bedömas ha låg risk för bias totalt sett, ska risken ha bedömts som låg i samtliga riskområden.

- För att utfallet ska bedömas ha hög risk för bias totalt sett ska risken vara hög i minst ett riskområde eller att studien har måttlig risk i flera riskområden.

NRSI-studier kan också ha en oacceptabelt hög risk, då minst en av de tre första riskområdena bedöms ha oacceptabelt hög risk för bias. Utfallet från sådana studier ska inte ingå i det fortsatta arbetet.

6.1.3 Steg 3: Jäv/Intressekonflikter

Frågan om jäv förekommer i studien besvaras enklast med hjälp av någon som har kännedom om det aktuella forskningsområdet. Uppgifter om jäv rapporteras förslagsvis i anslutning till beskrivning av risk för bias.

6.1.4 Steg 4: Sammanställning av total risk för bias per utfallsmått för alla studier

Det kan vara bra att sammanställa bedömningarna av risk för bias för alla studier. Ett sätt är att göra en tabell över risk för bias, exempelvis i excel, där man markerar studiernas risk för bias för olika riskområden. Exempelvis kan ett grönt fält innebära att studien har låg risk för bias inom det riskområdet, en grå markering innebär måttlig risk för bias för SBU, medan ett rött fält slutligen innebär hög risk för bias. Med hjälp av exempelvis Revman, ett verktygsprogram för systematiska översikter framtaget av Cochrane Collaboration, eller verktyget robvis, kan man också ta fram en sådan tabell. Observera att i tabellen som man tar fram i Revman klassificerar Cochrane mellannivån för risk of bias som ett gul-markerat riskområde, och benämner det som ”oklar risk” för bias, vilket är en terminologi som inte används av SBU. Ett exempel på en risk-för-bias-tabell visas i Figur 6.2.

Figur 6.2 Exempel på risk för bias-tabell. Ett grönt plus innebär att studien har låg risk för bias inom det riskområdet, medan en grå cirkel innebär måttlig risk för bias, och ett orange minus innebär hög risk för bias.

	Studie A 2017	Studie B 2018	Studie C 2018	Studie D 2019	Studie E 2019	Studie F 2020
Fördelningsfel (selection bias)	+	+	+	+	+	+
Behandlingsfel (performance bias)	+	+	−	+	+	−
Bedömingsfel (detection bias)	+	+	○	+	+	○
Bortfallsfel (attrition bias)	−	○	−	+	+	−
Rapporteringsfel (reporting bias)	−	+	+	○	+	−

6.2 Risk för bias i studier om diagnostisk tillförlitlighet

Sensitivitet och specificitet påverkas av olika typer av bias. Några av dem överensstämmer med dem som finns för andra studietyper, till exempel bias som

uppstår när den som tolkar resultaten inte är blindad. Andra är specifika för diagnostisk tillförlitlighet. Det finns en systematisk översikt som undersökt hur sensitivitet och specificitet påverkas av olika typer av bias [110]. Resultatet har sammanfattats i en tabell som nås här.

Studier om diagnostisk tillförlitlighet bedöms med stöd av granskningsmallen Quality Assessment of Diagnostic Accuracy Studies version 2, QUADAS-2 [111]. QUADAS-2 är i första hand utvecklad för att bedöma tvärsnittsstudier. Den är inte avsedd för att bedöma studier om prognos. Den engelskspråkiga versionen med sina detaljerade instruktioner finns på webbplatsen för Bristols universitet.

QUADAS-2 består av fyra riskområden med stödfrågor (eng. signalling questions): population, indextest, referensstandard samt tid och (process)flöde. Liksom för övriga granskningsmallar bedöms först risken för varje enskilt riskområde och sedan görs en sammanvägd bedömning av den totala risken. Till skillnad från övriga granskningsmallar tar QUADAS-2 upp såväl risk för bias som aspekter på tillämpbarhet (dvs. variation i jämförelse med projektets forskningsfråga) under varje riskområde (se Figur 6.3).

Figur 6.3 QUADAS-2 består av fyra riskområden. Först bedöms risken för bias för varje område, därefter görs en sammanvägd bedömning av den totala risken.

Innan granskningen påbörjas ska projektgruppen avgöra om QUADAS-2 behöver anpassas. Några signalfrågor kan vara överflödiga beroende på förutsättningarna för översikten och andra kan behöva läggas till. Man bör undvika att lägga till för många signalfrågor. En annan del av anpassningen är att besluta hur lång tid mellan indextest och referenstest som är acceptabel. När projektgruppen kommit överens om vilka signalfrågor som ska ingå bör granskningsmallen testas på ett mindre antal studier.

Första delen av QUADAS-2 är att rita upp ett flödesschema för hur en studie har genomförts. Det underlättar den fortsatta granskningen.

6.2.1 Riskområden i QUADAS-2

1. Population

1 a) Risk för bias

I idealfallet ska en studie rekrytera ett konsekutivt eller randomiserat urval av deltagare med olika risk för att de har tillståndet eller problemet ifråga, det vill säga ett brett spektrum av deltagare. Om ett smalt spektrum av deltagare ingår finns det risk för att sensitiviteten överskattas, så kallad spektrumbias [110]. Fall-kontrollstudier ska undvikas eftersom de bara inkluderar deltagare som antingen har tillståndet eller inte har tillståndet. Spektrumbias uppstår även i studier med randomiserat eller konsekutivt urval om vissa deltagare systematiskt utesluts, vilket kan leda till såväl över- som underskattning.

1 b) Tillämpbarhet

Det kan finnas flera orsaker till bristande tillämpbarhet. Exempel på sådana är i vilket skede av den diagnostiska processen som testet är tänkt att användas och om deltagarna är mer eller mindre selekterade. Demografiska skillnader och svårighetsgraden av tillståndet eller problemet kan också påverka tillämpbarheten, liksom skillnader i prevalens. En högre prevalens ökar sensitiviteten och minskar specificiteten [110].

2. Index test

2 a) Risk för bias

Detta riskområde avser två aspekter, blindning och val av tröskelvärde. Om indextestet genomförs efter referenstestet kan kännedom om referenstestets resultat påverka tolkningen av indextestet. Många tester har tröskelvärden, som kan vara mer eller mindre etablerade. En del studier kan ha valt att inte definiera tröskelvärdet i förväg utan väljer tröskelvärdet efteråt för att optimera testets prestanda, ett så kallat datadrivet tröskelvärde. Detta leder till risk för bias.

2 b) Tillämpbarhet

Om testet genomförs eller tolkas på ett annorlunda sätt än vad som avsågs i forskningsfrågan kan tillämpbarheten minska. Resultaten kan till exempel i studien tolkas av specialister, medan testet i praktiken är tänkt att användas av personer med mindre kunskap och erfarenhet. Olika versioner av ett test kan också bli ett problem.

3. Referensstandard

3 a) Risk för bias

Referensstandarden kan ge upphov till risk för bias. Referensstandarden förutsätts klassificera tillståndet eller problemet med 100-procentig korrekthet.

Tillförlitligheten kan dock påverkas om referensstandarden genomförts eller tolkats på ett bristfälligt sätt. Sådan så kallad felklassifikationsbias leder vanligen till att sensitiviteten överskattas [110].

Om indextestet genomförts före referensstandarden kan det också finnas risk för att en oblindad tolkare av referensstandarden blir påverkad av resultatet av indextestet.

3 b) Tillämpbarhet

Frågan om tillämpbarhet gäller främst om tillståndet eller problemet är definierat på samma sätt i studien som i projektets frågeställning (PIRO).

4. Tid och flöde

Om det går tid mellan testerna kan det finnas risk för att tillståndet eller problemet hunnit förändras till det bättre eller sämre, det vill säga att det blir en felklassificering. Risken för att en fördröjning mellan tester påverkar tillförlitligheten varierar mellan olika tillstånd och problem. Några dagars fördröjning spelar till exempel en mindre roll vid en kronisk sjukdom än vid akuta infektioner. Ett problem i sammanhanget är att vissa referensstandarder kan mätas först efter en längre tid. Ett exempel är när referensstandarden är en sjukdom och samtliga diagnostiska kriterier måste vara uppfyllda.

När det kommer till flöde kan verifikationsbias uppstå. Det innebär att endast en del av deltagarna undersöks med den optimala referensstandarden. Övriga deltagare bedöms inte med någon referensmetod alls (partiell verifikationsbias), eller så väljs en annan, enklare, referensstandard (differentiell verifikationsbias). Om resultatet av indextestet påverkar valet av referensstandard uppstår systematisk bias. Orsaker till ett sådant beslut kan vara att referensstandarden till exempel är dyr eller att undersökningen kan medföra risker för deltagaren.

En andra aspekt av flöde rör bortfallet. Om inte alla som rekryterats finns med i analysen uppstår bias eftersom sådana som fallit bort tenderar att skilja sig systematiskt från dem som är kvar.

6.3 Bedömning av studier med kvalitativ metodik

Flera begrepp har utvecklats för att beskriva vilka mått och steg som forskarna vidtagit för att öka resultatens tillförlitlighet, beroende på forskningstradition [26]. Ett av dem bygger på den kvantitativa traditionen och bedömer validitet, reliabilitet och generaliserbarhet. Ett mer använt begrepp är trovärdighet (eng. trustworthiness) [112] som består av fyra komponenter (se Faktaruta 6.1). Ett tredje begrepp är vetenskaplig stringens (eng. scientific rigour), ett begrepp som används i Cochranes handbok.

Faktaruta 6.1 Komponenter i begreppet trovärdighet i kvalitativa studier [113].

Credibility: I vilken utsträckning data och analytisk process adresserar fokus för studien, om det finns tillräckliga data och om avvikande data hanterats på ett adekvat sätt.

Transferability: Forskarna kan visa att resultaten är användbara i likartade sammanhang och för likartade frågor. Här behövs detaljerade beskrivningar av sammanhangen (kontexten) som stöd för bedömningen.

Dependability: Visar om förutsättningar och sammanhang är lika över tid.

Confirmability: Visar att resultaten bekräftats av andra och att de tycker att tolkningarna är rimliga.

Forskare har olika åsikter om huruvida det har något värde att bedöma metodbrister i en studie som har använt kvalitativ metodik [114]. SBU anser att det är viktigt att en syntes av forskningen bygger på tillförlitliga studier och de kriterier som används för att bedöma tillförlitligheten i kvalitativa studier överensstämmer i huvudsak med dem som används för kvantitativa studier.

Studiens forskningsfråga ska bäst besvaras med en kvalitativ metod och valet av ansats ska motiveras. 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. Tillförlitligheten ökar om tolkningen har verifierats, exempelvis genom att flera forskare analyserar materialet oberoende av varandra eller genom att preliminära tolkningar diskuteras med utomstående [26].

Idag finns drygt 100 publicerade checklistor som stöd för att identifiera brister i genomförande och rapportering av kvalitativa studier [107] där Critical Appraisal Skills Programme (CASP) är ett av de mer etablerade. Ingen av checklistorna stödjer en bedömning av risken för att identifierade brister påverkar fyndens tillförlitlighet, något som även noterats av Cochrane Collaboration [115]. Cochrane utvecklar därför ett eget formulär. Tillsvidare rekommenderar Cochrane att utgå från vissa frågor i CASP.

SBU har valt att istället utveckla en egen granskningsmall. Den är, i likhet med övriga granskningsmallar som används av SBU, uppbyggd av olika områden med tillhörande stödfrågor. Fokus ligger på att bedöma risk för att metodbrister påverkar resultaten. Men, till skillnad från de övriga granskningsmallarna finns även frågor som ska underlätta den kommande bedömningen av tillförlitligheten i det syntetiserade resultatet med stöd av CERQual. Mer detaljerad vägledning till hur mallen fungerar finns i dess instruktioner.

6.3.1 Metodbrister

Metodbrister granskas utifrån fem aspekter:

1. överensstämmelse mellan teoretisk underbyggnad av studien och dess syfte
2. urval
3. datainsamling
4. analys

5. forskarens roll.

Varje aspekt ovan består i sin tur av tre moment:

1. Gör en kort beskrivning av till exempel urvalsprinciper eller vilka metoder som användes för att samla in data. Syftet med denna del av mallen är att underlätta att skriva rapportmanus.
2. Besvara frågorna som ska stödja bedömningen.
3. Överväg de brister som identifierats och i vilken utsträckning det finns risk för att de påverkar fynden. Det finns tre fasta bedömningsalternativ: **1)** Ja det finns en allvarlig risk; **2)** Nej, risken bedöms inte vara allvarlig och **3)** Oklart, det finns inte tillräcklig information för att bedöma risken. Dessa bedömningar läggs sedan in i en sammanställning som används i GRADE-CERQual.

Den sista delen av granskningen är att bedöma om studien sammanlagt har så allvarliga problem att den inte bör ingå i metasyntesen. Observera att ett problem också kan vara att studien är så klent beskriven att det inte går att bedöma riskerna.

För studier som bedöms ha låg eller måttlig risk för att resultaten påverkats av metodbrister fortsätter man att besvara de övriga frågorna i granskningsmallen.

6.3.2 Överförbarhet

Den kvalitativa forskningens resultat är ofta beroende av sammanhanget och läsaren måste därför noggrant bedöma i vilken utsträckning de är överförbara till andra sammanhang, graden av transferabilitet. Bedömningen underlättas om författarna diskuterar 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.

Överförbarhet till forskningsfrågans SPICE (eller motsvarande) hanteras som relevans i GRADE-CERQual (se avsnitt 10.2). Brister i överförbarheten för en enskild studie såväl som för hela underlaget klassificeras som indirekt, partiell eller osäker.

En indirekt överförbarhet uppkommer då någon del av inklusionskriterierna, till exempel perspektiv eller setting, har bytts ut. Ett exempel på detta såg man när man tittade på faktorer som påverkde implementeringen av vaccinationskampanjer för fågelinfluensa [116]. Här saknades studier om fågelinfluensa men det fanns studier om svininfluensa. En förutsättning för att använda sig av underlag med indirekt relevans är att det finns tillräckligt med gemensamma faktorer.

Relevansen är partiell om studierna inte täcker hela forskningsfrågan, till exempel

om alla studier är genomförda i ett land. En osäker relevans uppstår om studierna har knapphändig information om till exempel deltagare eller sammanhang.

6.3.3 Koherens

Med koherens avses i vilken utsträckning resultaten bygger på alla ingående data. Har de tagit hänsyn till data som inte passar in i mönstret? Om det finns mycket material som inte ingår, har forskarna då sökt alternativa förklaringar som bättre täcker data? Här finns svarsalternativen ”ja”, ”nej” och ”oklart”.

6.3.4 Tillräckliga data

Här bedöms kvantitet och kvalitet på data som studien bidrar med. En aspekt är om antalet deltagare är tillräckligt. Det finns inga regler för hur stort urvalet ska vara inom kvalitativ forskningsmetodik. Antalet informanter avgörs dels av syftet med studien, dels av egenskaper hos deltagarna och forskaren. Ofta används begreppet datamättnad (eng. saturation) för den gräns när forskarna bedömer att ytterligare datainsamling inte ger mer kunskap. Datamättnad kommer från grounded theory men används även i andra sammanhang. En studie med en avgränsad fråga, eller som har djupintervjuer med informanter som bidrar med rika data, kräver färre deltagare än studier med bredare frågor och med metoder för datainsamling som ger mindre information. Ytterligare en faktor som påverkar urvalsstorleken är hur förtrogen datainsamlaren är med ämnesområdet och insamlingstekniken [117].

Den andra aspekten är om data är detaljerat, det vill säga tjockt och rikt, eller mer översiktligt och därmed ytligt (”tunt”). Högt strukturerade intervjuformulär och användning av fokusgrupper kan vara några anledningar till att data är tunt.

6.4 Granskning av systematiska översikter med ROBIS

SBU använder för närvarande två granskningsmallar för granskning av systematiska översikter: en anpassad version av AMSTAR (A measurement tool to assess systematic reviews) [118] och ROBIS (Risk of bias in systematic reviews) [119]. Valet av granskningsmall beror på vilken rapporttyp översikten ska ingå i. Båda granskningsmallarna bottnar i PRISMA-riktlinjer för systematiska översikter och utvärderar hur de olika delarna av processen genomförts. Den anpassade versionen av AMSTAR stödjer en snabb granskning av stora mängder systematiska översikter. Detta beskrivs närmare i Kapitel 11. Styrkan med ROBIS är att den fokuserar på översiktens risk för bias. Därmed kan bedömningen lätt integreras i GRADE (se Kapitel 9).

ROBIS har utvecklats av Cochrane Collaboration. Originalformuläret på engelska med en detaljerad manual finns här. Den som inte har tidigare erfarenhet av att bedöma systematiska översikter med ROBIS rekommenderas

att läsa igenom manualen. SBU har översatt formuläret till svenska och lagt till en fråga om intressekonflikter.

6.4.1 Struktur på ROBIS

ROBIS är uppbyggd på samma sätt som mallarna för primärstudier, med riskområden och stödfrågor. Bedömningen görs i tre steg:

1. Relevans. På SBU är relevansbedömningen oftast redan genomförd innan översikten granskas med ROBIS.
2. Identifiering av eventuella brister i översiktens arbetsprocess. Formuläret är indelat i fyra riskområden: kriterier för val av studier (PICO), identifiering och val av studier, datainsamling och bedömning av studierna och de ingående utfallens risk för bias samt analys och slutsatser.
3. Bedömning av den sammantagna risken för bias görs med stöd av fyra frågor. Svarsalternativen är "låg", "hög" eller "oklar risk".

I detta kapitel beskriver vi översiktligt de olika riskområdena i ROBIS. Mera detaljerad information finns i SBU:s steg-för-steg-instruktioner.

6.4.2 Riskområde 1: Kan urvalskriterierna leda till risk för bias?

Grunden är att en systematisk översikt ska ha en fördefinierad forskningsfråga med specificerade inklusionskriterier. För att kunna bedöma om författarna har gjort avsteg från frågan eller inklusionskriterierna behövs tillgång till översiktens protokoll eller forskningsplan. Förr publicerades sällan protokollen, med undantag för Cochrane Collaborations systematiska översikter, men numera registreras ofta protokollen i databasen PROSPERO och många författare publicerar också protokollet som en vetenskaplig artikel.

6.4.3 Riskområde 2: Leder brister i sökningen till att relevanta studier missats?

En otillräcklig sökstrategi kan leda till att relevanta studier inte kommer med i sökningen och att översiktens resultat blir otillförlitliga. Det kan vara svårt att avgöra om en sökstrategi är acceptabel. Ett sätt att få en kvalitetskontroll i bedömningen av översikten är att projektledare eller sakkunniga först gör en preliminär bedömning. Om de bedömer att sökstrategin är bristfällig exkluderas studien. Om det är svårt att bedöma sökstrategin men översikten kan anses ha hög tillförlitlighet utifrån bedömningen av de övriga riskområdena kontrollerar en informationsspecialist sökstrategin. Först därefter görs en slutlig bedömning.

6.4.4 Riskområde 3: Påverkas resultaten av brister vid bedömning av studier?

Riskområdet tar upp två aspekter: om processen är utförd enligt PRISMA:s riktlinjer och hur författarna har bedömt riskerna för bias i de ingående studierna. Äldre översikter har inte bedömt risken för bias utan artiklarnas studiekvalitet med till exempel Jadads kriterier [120] eller SIGN, som har

utvecklats av Scottish Intercollegiate Guidelines Network [121]. Ofta har författarna summerat hur många kriterier som uppfyllts och satt ett tröskelvärde för när studierna kan anses ha måttlig eller hög kvalitet. Sådana poäng är inte lätta att överföra till risker för bias för ett enskilt utfallsmått och det kan finnas viktiga aspekter som överhuvudtaget inte tas upp.

6.4.5 Riskområde 4: Påverkas resultaten av brister i analysen?

Riskområdet omfattar en bedömning av om författarna har använt en bra metod för att analysera sina resultat och om det finns risk för publikationsbias.

6.4.5.1 Är analysmetoden adekvat?

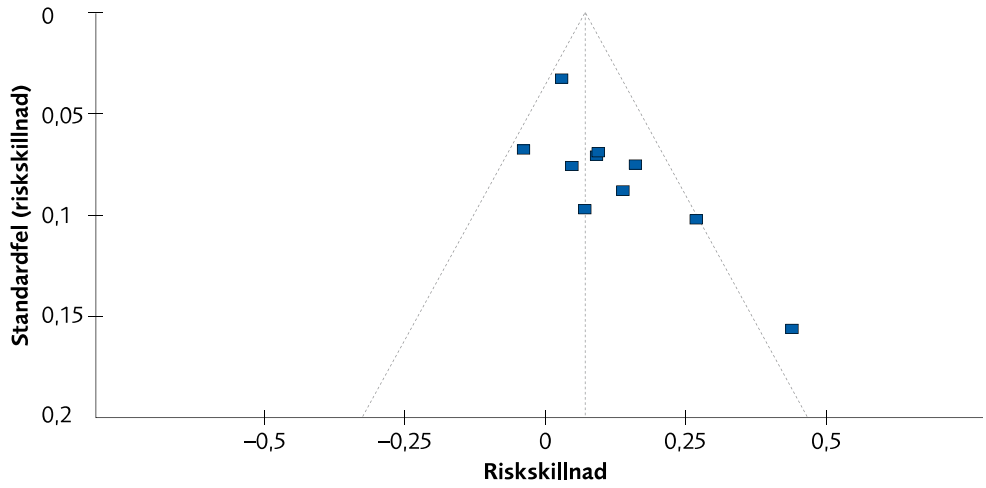
Ett första övervägande gäller om det finns metaanalyser och om det var lämpligt att göra en metaanalys. Är studierna tillräckligt homogena? Författarna bör också motivera sina val av metod för metaanalys och valet ska vara tillfredsställande (se Kapitel 8 för närmare beskrivning av olika metoder).

6.4.5.2 Publikationsbias

Med publikationsbias avses att studier av olika skäl inte publiceras alls eller med tidsfördröjning [122]. Den vanligaste orsaken är att studien inte kunnat finna några positiva, signifikanta resultat, vilket kan göra såväl forskaren, som en eventuell sponsor, samt tidskrifter mindre benägna att publicera studien. Om studien är stor ökar chansen att den publiceras, men med en fördröjning på nära två år jämfört med om den haft positiva resultat (eng. lag time) [123]. Det finns följaktligen en risk att metaanalysens resultat snedvridits (oftast överskattats) på grund av att opublicerade studier inte finns med. Det är ofta mycket svårt att fastställa om det råder publikationsbias men det finns verktyg som stöd för en bedömning [124] [125]. Översiktens författare bör ha försökt bedöma risken och redovisa resultatet av den.

En vanlig metod för att få en uppfattning om risken är att göra ett så kallat trattdiagram (eng. funnel plot) (se Figur 6.4) [124]. Trattdiagrammet kan konstrueras i RevMan. Förutsättningen är att det finns många publicerade studier, ofta nämns ett minimum på 12 studier. I diagrammet läggs storlek och resultat från varje studie in. Om det inte finns någon publikationsbias liknar resultatet en symmetrisk upp-och-nervänd tratt (därav namnet). Om grafen är asymmetrisk, framförallt att små studier med negativa resultat saknas, kan det finnas skäl att misstänka publikationsbias. Det kan dock finnas andra orsaker bakom asymmetrin, så att enbart använda trattdiagram räcker inte för att påvisa publikationsbias [124].

Figur 6.4 Exempel på trattdiagram (funnel plot). Varje studie representeras av en punkt. Den horisontella axeln visar effektstorleken medan den vertikala visar spridningen (standardfelet). Ju högre upp på axeln ett resultat ligger, desto mindre är spridningen.



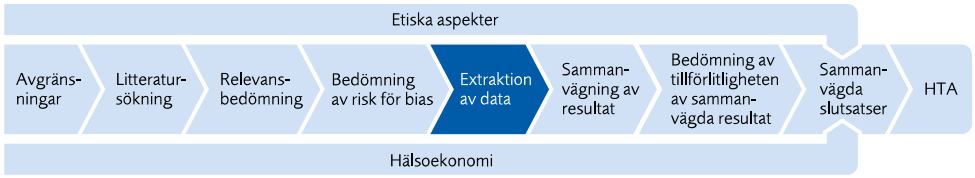
Metoder för att analysera eventuell publikationsbias

För att undersöka publikations bias ytterligare kan man till exempel genomföra Egger´s test. Man kan också välja att spegla studier som ligger långt ut och se vad som skulle hända med det gemensamma effektmåttet om det fanns motsvarande studier på andra sidan av mittlinjen i trattdiagrammet. Dessa åtgärder kan inte genomföras i RevMan, men finns tillgängliga i programmet CMA.

6.5 Granskning av systematiska översikter av kvalitativ forskning

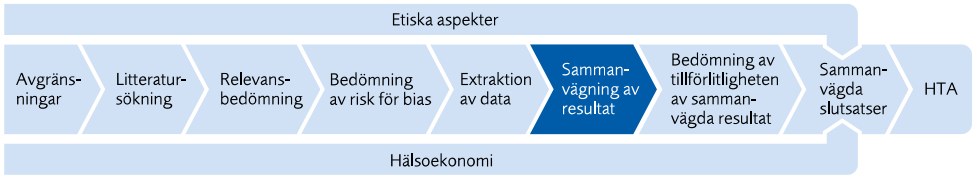
Det finns inga publicerade mallar eller checklistor som stöd för att bedöma risk för att fynden påverkats av metodproblem i systematiska översikter av kvalitativ forskning. Många aspekter är desamma som för systematiska översikter av kvantitativ forskning men några skiljer sig åt. SBU har därför tagit fram en mall som bygger på dels ROBIS (se avsnitt 6.4) och dels ENTREQ-riktlinjerna för att genomföra och rapportera kvalitativa översikter [3].

7. Extraktion av data



Nästa steg i att göra en systematisk översikt är att extrahera data från de ingående studierna och lägga in dem i tabeller. Syftet är att de som läser översikten ska få en uppfattning om karakteristika för studierna utan att ha läst dem själva. Typiska uppgifter som ska finnas i tabellerna är författare, beskrivning av populationen och den miljö som studien genomförs i, beskrivning av deltagarna i studien (ålder, kön etc.), beskrivning av intervention och kontroll (eller indextest och referensstandard etc.) samt utfallsmått som använts i studien. Tabellerna sammanställs på engelska.

8. Sammanvägning av resultat



Nästa steg är att undersöka resultaten från de studier som utgör det vetenskapliga underlaget och bedöma om det går att dra några slutsatser om till exempel effekten av en intervention.

För studier som bygger på kvantitativ metodik, till exempel effekter av interventioner eller diagnostisk tillförlitlighet, är det önskvärt att använda metaanalys för att väga samman resultaten. Beroende på analysmodell ger metaanalysen antingen en uppskattning av en antaget gemensam underliggande effekt (eller sensitivitet och specificitet) eller ett medelvärde av effekterna (eller sensitiviteten och specificiteten). När det av olika skäl inte går att göra en metaanalys får man istället beskriva resultaten, det vill säga att man gör en narrativ syntes.

För studier som har använt kvalitativ metodik finns flera metoder för att syntetisera fynden.

Denna del av metodboken beskriver de olika tillvägagångssätten för sammanvägning som används av SBU: metaanalys, narrativ sammanställning samt några alternativ för syntes av kvalitativa fynd.

8.1 Metaanalys för interventionsstudier

Detta är en översiktlig beskrivning av metoden och mer detaljerad information finns till exempel i boken Introduction to Meta-analysis av Borenstein och medarbetare [126] eller i Cochranes handbok för systematiska översikter [47].

Metaanalysen utvecklades som en hjälp för att få fram mera pålitliga resultat genom att data från flera enskilda studier läggs samman med hjälp av statistiska metoder (eng. pooling). Det gemensamma, sammanvägda resultatet uttrycks sedan som ett punktestimat med ett tillhörande osäkerhetsintervall (konfidensintervallet).

Eftersom en metaanalys består av flera studier, och innehåller mer data än en enskild studie så leder det till en ökad så kallad teststyrka (eng. statistic power). En ökad teststyrka ger bättre möjligheter att upptäcka effekter som faktiskt finns, såsom skillnader mellan en interventions- och en kontrollgrupp. Den ökade teststyrkan och den större mängden individer och händelser som ingår i metaanalysen gör också att skattningen på det effektmått som räknas fram troligen ligger närmare det sanna värdet för den bakomliggande populationen, och inte bara för de individer som ingår i en viss studie, stickprovet (se Faktaruta

8.1).

Ibland kan de studier som finns att tillgå vara för olika varandra för att det ska vara meningsfullt att beräkna ett sammantaget estimat. Men även när en metaanalys inte kan användas för att beräkna punktestimatet kan tekniken ge värdefull information. Metaanalysen kan användas som ett verktyg för att analysera olika källor till variation i materialet (t.ex. urvalsfel och heterogenitet), och för att undersöka risken för publikationsbias i det vetenskapliga underlaget (se Kapitel 9).

Faktaruta 8.1 Skillnad mellan population och stickprov.

Population: Alla som är av intresse (till exempel alla patienter i Sverige som lider av lungcancer).

Stickprov (eng. sample): Den del av populationen som ingår i en studie. Stickprovet ska vara representativt för populationen, så att slutsatser som dras från stickprovets data kan generaliseras till att gälla för hela populationen. Stora och helst randomiserade stickprov ökar tillförlitligheten vid generaliseringar. Metaanalyser baseras på ett större stickprov än enskilda studier, eftersom de innehåller fler individer och mer data.

8.1.1 Utfallsmått i en metaanalys

En metaanalys gäller ett specifikt utfall som mätts på ett specifikt sätt. Ofta har dock studierna mätt utfallet på olika sätt, det vill säga informationen finns men är i fel format. Resultaten måste då räknas om för att kunna användas i analysen.

Utfallsmått kan klassificeras som kategoriska eller kontinuerliga. Kategoriska mått hanterar ett begränsat antal nivåer, till exempel olika utbildningsnivåer (grundskola, gymnasium, kandidatexamen, mastersexamen, doktorsexamen). Ett specialfall är dikotoma (binära) mått som hanterar händelser som kan översättas till ettor och nollor. Antingen har en händelse inträffat eller också har den inte det. Kontinuerliga variabler hanterar mått som inte har några fasta nivåer, till exempel blodtryck, och uttrycks ofta som medelvärden eller medelvärdesskillnader.

För resultat som uttrycks med dikotoma eller kategoriska mått kan det sammanvägda resultatet, estimatet, uttryckas på flera sätt (se Faktaruta 8.2).

Faktaruta 8.2 Tre vanliga utfallsmått för kategoriska och dikotoma data.

	Händelse	Inte händelse	Totalt antal individer
Experimentgrupp	A	B	Tot exp (A+B)
Kontrollgrupp	C	D	Tot kont (C+D)

Riskskillnad (eng. risk difference, RD): Absolut mått på skillnader i risk.
RD= (A/Tot exp) – (C/Tot kont)

Relativ risk (eng. relative risk, RR): Relativt mått, man ska notera att även små riskdifferenser kan ge stora riskkvoter. RR kan inte beräknas i fall-kontrollstudier.
RR= (A/Tot exp) / (C/Tot kont)

Oddskvot (eng. odds ratio, OR): Är ett matematiskt stabilare mått än RR och kan till skillnad från RR beräknas även i fall-kontrollstudier.
OR= (A/B) /(C/D)

Oddskvoter kan även beräknas via logistisk regression om man vill väga in confounders i sin analys.

Oddskvoter – ett matematiskt mer stabilt mått

Oddskvot (eng. odds ratio, OR) är ett matematiskt stabilare mått än relativ risk (RR) och fungerar även vid fall-kontrollstudier då risken inte kan beräknas. Bland annat är det vändbart, så att du kan titta både på events jämfört med non-events, eller på komplementhändelsen non-events jämfört med events. Detta går inte att göra med RR. En annan fördel är att man kan beräkna OR med logistisk regression så att man lätt kan lägga till confounders till modellen.

Kontinuerliga mått mäts längs ett kontinuum och kommer i en jämförande analys att ge ett estimat uttryckt som medelvärdesskillnad, MD, eller standardiserad medelvärdesskillnad, SMD. Om alla ingående studier redovisar resultat från samma mätskala bör MD användas. I vissa fall går det att konvertera resultat från olika mätskalor till en enda och uttrycka resultatet som MD.

SMD används när resultaten bygger på mätning med olika skalor. En förutsättning är att mätskalorna mäter likartade egenskaper och det är en viktig bedömning vilka mätskalor som kan ingå i en metaanalys. Man kan behöva göra flera metaanalyser där man endast lägger ihop resultaten från de studier som använt samma skala eller mätt utfallet på samma sätt. Läs mer om detta nedan.

Metaanalys om samma skala används i studierna

Om studierna har använt samma skala kan det sammanslagna effektmåttet presenteras som medelvärdesskillnaden i originalskalans skalsteg. Fördelen är att man inte behöver göra några konverteringar. Läsaren måste dock ges tillräckligt med information för att kvalitativt kunna bedöma storleken på effekten. Det är en fördel om skalan är väletablerad och ofta använd inom forskningsområdet, eller intuitiv för läsaren att förstå. Skalan bör förklaras på så sätt att både dess minimum och maximum, samt betydelsen av negativa och positiva värden går att förstå. Har man tillgång till pålitliga bedömningar av dess minsta kliniskt betydelsefulla skillnad (eng. minimal important difference, MID, eller minimal clinically important difference, MCID) kan detta vara värdefull information i detta sammanhang.

Om skalan varken är välanvänd inom forskningsområdet eller intuitiv att förstå bör man vara särskilt noggrann med att beskriva skalan och överväga att även presentera effektstorleken i ett annat format (t.ex. dikotomiserat eller konverterat till en annan mer använd skala, se Tabell 8.1 i klickrutan "Metaanalyser med kontinuerliga utfallsmått och tolkning av SMD" nedan.

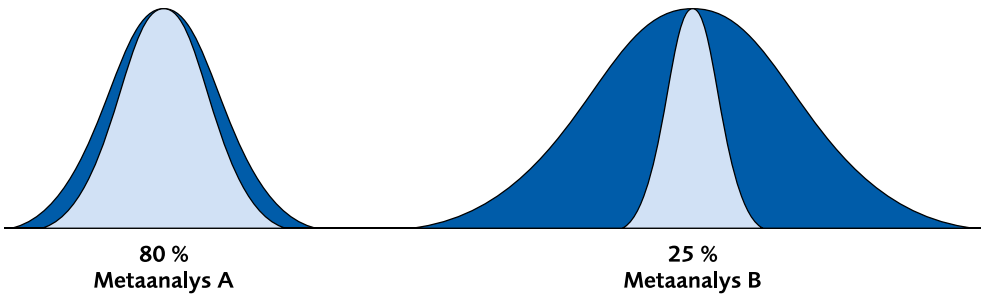
Om utfallen mätts på för olika sätt är inte metaanalys lämplig.

8.1.2 Heterogenitet

Olika studier skiljer sig oftast från varandra med avseende på upplägg, stickprovets eller studiepopulationens sammansättning, kontext (sammanhang), interventionernas exakta innehåll, kontrollvillkoren, sättet att mäta effekterna, studiedesign och annat. Detta leder till så kallad heterogenitet och innebär att de olika studierna kan såväl över- som underskatta effekten i den bakomliggande populationen. Studier som är för heterogena bör inte slås ihop i en metaanalys. Orsaken är att man då dels får en variation som beror på annat än det som vi undersöker, vilket kan dölja verkliga effekter, dels att man riskerar att dra felaktiga slutsatser från våra data.

Heterogenitet innebär alltså att det finns en variation i effektstorlek mellan studier, utöver den variation som vi förväntar oss att ska finnas, som beror på slumpen (variation inom studier). Metaanalysen ger oss tre mått för att undersöka heterogeniteten i ett material; **τ2**, **I2** och **Q**, där I2 är det mått som man oftast brukar använda för att bedöma heterogenitet. Måttet I2 ger en uppfattning om hur stor andel av den totala variationen i metaanalysen som beror på skillnader mellan de ingående studierna, och hur stor andel som beror på spridningen inom varje enskild studie (se Figur 8.1). Enligt en omtvistad tumregel sägs heterogeniteten vara låg om I2 ligger runt 0,25, måttlig om I2 ligger runt 0,50 och hög om I2 ligger runt 0,75, men att använda I2 som ett direkt mått på heterogenitet kan vara svårtolkat och rekommenderas inte. Måttet τ2 (Tau2) ger en uppfattning om hur stor skillnaden, i genomsnitt, är mellan punktestimaten för de olika studierna som ingår i analysen, medan måttet Q istället visar den genomsnittliga skillnaden mellan punktestimaten för de ingående studierna och metaanalysens gemensamma, sammanvägda estimat.

Figur 8.1 Den totala variationen som finns i en metaanalys (skillnader *inom* enskilda studier plus skillnader *mellan* enskilda studier) representeras här av den totala arean under den svarta kurvan. Den andel av variationen som beror på skillnader mellan de ingående studierna, I2, representeras av arean av det ljusblåa fältet. För metaanalys A utgör I2 80 procent av den totala variationen, medan I2 bara utgör 25 procent för metaanalys B. Den totala variationen i materialet är dock mycket större i metaanalys B.



Ofta är det omöjligt att få korrekta skattningar av heterogenitet. Om metaanalysen bygger på få studier finns det risk för att uppskattningen felaktigt visar att det inte finns någon mellanstudievarians, det vill säga ett falskt intryck av homogenitet. Rent generellt har heterogenitetstest en låg statistisk teststyrka

och heterogeniteten underskattas ofta i meta-analyser. τ2, I2 och Q kan dock vara användbara för att få en uppfattning om heterogeniteten i en metaanalys, och vara en grund till diskussion.

Det finns olika sätt att hantera att studier som man lägger ihop inte är helt lika, till exempel med hjälp av olika statistiska modeller eller med hjälp av subgruppsanalyser, se nedan.

8.1.3 Subgrupper i en metaanalys

Ett sätt att hantera heterogenitet mellan studier är att göra subgruppsanalyser. Sådana analyser ska vara planerade i förväg i projektplanen, och det ska finnas en tydlig orsak till att de är valda, till exempel att man har anledning att misstänka att kvinnor och män reagerar olika på en viss behandling. Att skapa subgrupper i efterhand, på basis av hur redan analyserade data ser ut, är inte att rekommendera. Det är också viktigt att överväga vilken teststyrka som man får i de olika subgrupperna. Eftersom varje subgruppsanalys utgör ett mindre stickprov än vad som skulle ha varit fallet utan subgrupperingar så kommer teststyrkan att minska, och därmed minskar också möjligheten att upptäcka eventuella skillnader som kan finnas i materialet.

Några exempel på tänkbara subgruppsanalyser är om studierna rapporterar olika varianter av interventionen, kommer från olika länder med olika sjukvårdssystem, har olika uppföljningstider eller när äldre studier använder en annan teknik än nyare studier.

8.1.4 Val av modell för metaanalys

Det finns två huvudtyper av metaanalyser, Fixed effect model (FEM) och Random effects model (REM). Vilken av modellerna som ska användas ska bestämmas redan i projektplanen och beror på vilket syfte översikten har.

FEM utgår från antagandet att alla studier som ingår i metaanalysen är stickprov som har dragits från en och samma population. Därmed tänker man sig att det finns en gemensam effekt som gäller för den bakomliggande population som alla studierna har dragits ifrån. Det är denna gemensamma effekt som man vill skatta i metaanalysen. Till exempel kanske man antar att alla patienter i Norden som behandlas för närsynthet är en tillräckligt homogen grupp för att man ska kunna anta att de tillhör samma population oavsett om studierna kommer från Sverige, Norge eller Danmark, och man kan då välja att använda en FEM.

För REM antar man att de ingående studierna har dragits från olika bakomliggande populationer. Man kanske tycker att behandlingen mot närsynthet skiljer sig så pass mycket mellan de olika länderna att utfallet kan påverkas. Det går då inte längre att anta att det finns en gemensam effekt, eftersom effekten kan förväntas vara olika i de olika studiepopulationerna. REM ger istället en skattning av medelvärdet över alla olika ingående populationer. Man får därmed inte ett direkt mått på hur effekten ser ut i en enskild population (till exempel närsynthet i just Sverige). Å andra sidan kan skattningen

ge en uppfattning om var effekten ligger mer generellt, i genomsnitt.

Eftersom varje studie i REM blir den enda representanten för just sin population så får små avvikande studier större vikt än vid FEM, och konfidensintervallet blir bredare. En annan konsekvens är att den statistiska teststyrkan blir lägre för REM jämfört med FEM, och möjligheten att upptäcka faktiska skillnader mellan grupperna minskar något. Ju mindre heterogenitet som finns i en analys, desto mer lika blir dock de två modellerna. Ett problem med REM är att det är svårt att veta om man har täckt in alla möjliga populationer tillräckligt bra för att kunna få fram ett relevant medelvärde.

8.1.5 Tolkning av resultat

Metaanalys kan användas på flera olika sätt och den är ett bra analysverktyg för att få en bättre förståelse för de data man arbetar med. Vilka slutsatser som kan dras från resultatet av en metaanalys beror på hur lika studierna som ingår i analysen är. Översiktligt kan man säga att det finns tre nivåer, se Faktaruta 8.3.

Faktaruta 8.3 Tre nivåer av hur lika studierna som ingår i en metaanalys är.

1. De studier som ingår är i allt väsentligt lika.
   - Effekten kan anses vara robust för de studerade populationerna.
2. De studier som ingår skiljer sig åt, men på slumpartade vis.
   - Rapportera medeleffekten och diskutera vad spridningen av data kan bero på och vad det har för betydelse.
3. De studier som ingår skiljer sig åt vad gäller viktiga aspekter.
   - Medeleffekten är inte relevant. Diskutera spridningen av data och vad det kan bero på. Här sammanställs resultaten narrativt.

8.1.6 Tolkning av utfallet uttryckt som SMD

En nackdel med att använda SMD är att det kan vara svårt att kvalitativt uppskatta effektens storlek när den beskrivs som ett visst antal standardavvikelser. Det kan därför vara värdefullt att komplettera ett resultat i SMD med ett eller flera alternativa effektmått för att underlätta tolkningen av effektens storlek.

En annan nackdel med SMD (eller någon annan metod som använder SMD som mellansteg) är att resultatet är känsligt för faktorer som påverkar standardavvikelsen, till exempel urvalet och antalet personer i varje grupp. I sådana fall kan man med fördel även presentera resultatet på ett sätt som inte är beroende av standardavvikelsen.

För projekt med kontinuerliga utfallsmått rekommenderas läsning av en utförlig vägledning, inklusive kalkylatorer för olika sätt att omvandla SMD, och komplement till SMD. Vägledningen kan du läsa nedan.

Vägledning: Metaanalyser med kontinuerliga utfallsmått och tolkning av SMD

Definition

Denna vägledning gäller kontinuerliga utfallsmått, som kan antas vara normalfördelade och som man kan analysera med parametriska metoder, oavsett vilka värden variabeln kan anta inom sitt variationsområde.

Utfallsmått som inte kan antas vara normalfördelade kan man inte analysera med parametriska metoder. Detta gäller även för metaanalyser. Det finns icke-parametriska sätt att göra metaanalyser på men det är ovanligt. Utfallsmått som inte är normalfördelade kan man istället gruppera till endast två utfall. Därefter analyserar man dem som dikotoma utfall i en parametrisk metaanalys.

Kontinuerliga jämfört med dikotoma utfallsmått

Fenomen som är dikotoma till sin natur, alltså sådana fenomen som bara har två alternativ (t.ex. död, ej död) mäts med dikotoma utfallsmått. För fenomen som är kontinuerliga till sin natur, det vill säga att de i princip kan anta vilket värde som helst inom ett visst område (t.ex. data från blodprovsanalyser), är det inte lika enkelt. Dessa kan kategoriseras som både kontinuerliga och dikotoma utfallsmått. Oavsett vilket man väljer förlorar man viktig information när man aggregerar data från individ till gruppnivå.

Ett exempel är depression. Tillståndet kan mätas med ett kontinuerligt utfallsmått med hjälp av skattningsskalor. Det kan även kategoriseras i flera nivåer, till exempelvis mild, måttlig eller svår depression. Ytterligare ett alternativ är att dikotomisera utfallet, det vill säga deprimerad eller inte deprimerad. Om man väljer att dikotomisera depression går det att utvärdera hur många individer som inte längre var deprimerade efter behandlingen, jämfört med innan eller jämfört med en kontrollgrupp. I den beräkningen går man dock miste om exakt hur mycket mindre deprimerade individerna blev. Om man å andra sidan väljer att mäta depression med en skattningsskala går det att utvärdera hur mycket mindre deprimerade individerna var efter behandlingen. I den beräkningen går man å sin sida miste om hur många individer som uppnådde en viss minskning.

Båda perspektiven är dock relevanta och det är därför värdefullt att beräkna resultaten på båda dessa sätt om det är möjligt., oavsett vilken typ av utfallsmått man valt. För dikotoma utfallsmått kan det röra sig om att tydligt beskriva tröskeln för händelse/icke-händelse och för kontinuerliga utfallsmått att sätta resultatet i relation till hur stor andel av individerna som kan uppskattas uppnå en viss effekt. Om man inte har någon tröskel att relatera resultaten till så kan det hjälpa att påminna sig om att hälften av individerna estimeras få en effekt som är större än medelvärdet, och hälften en effekt som är mindre än medelvärdet (om man antar att normalfördelning råder).

1. Metaanalys om samma skala används i studierna

Om studierna har använt samma skala kan det sammanslagna effektmåttet presenteras som medelvärdesskillnaden i originalskalans skalsteg. Fördelen är att man inte behöver göra några konverteringar. Läsaren måste dock ges tillräckligt med information för att kvalitativt kunna bedöma storleken på effekten. Det är en fördel om skalan är väletablerad och ofta använd inom forskningsområdet, eller intuitiv för läsaren att förstå. Skalan bör förklaras på så sätt att både dess minimum och maximum, samt betydelsen av negativa och positiva värden går att förstå. Har man tillgång till pålitliga bedömningar av dess minsta kliniskt betydelsefulla skillnad (eng. minimal important difference, MID, eller minimal clinically important difference, MCID) kan detta vara värdefull information i detta sammanhang.

Om skalan varken är välanvänd inom forskningsområdet eller intuitiv att förstå bör man vara särskilt noggrann med att beskriva skalan och överväga att även presentera effektstorleken i ett annat format (t.ex. dikotomiserat eller konverterat till en annan mer använd skala, se Tabell 8.1 nedan för alternativa format).

2. Metaanalys om studierna har använt olika skalor

Steg 1: Bedöm om skalorna är tillräckligt lika för en sammanslagning.
Om studierna har använt olika skalor måste man först bedöma om skalorna mäter samma så kallade konstrukt eller inte. Ett konstrukt är ett konstruerat begrepp som kategoriserar vissa typer av fenomen. Som ett exempel så grupperar konstrukten frukt, möbel och verktyg vissa typer av saker. Det finns egentligen ingen frukt, möbel eller verktyg i sig utan begreppet benämner en konstruerad kategori. Det samma gäller även vissa utfallsmått i forskningsstudier, såsom stress, depression eller intelligens, som är grupperingar av flera enskilda fenomen.

Då man beslutar om man ska slå samman studier som använt sig av olika skalor vad gäller utfall ska man alltså bedöma om de skalor som använts i studierna mäter samma konstrukt, det vill säga rimligt likartade fenomen eller egenskaper. Skalor som delvis mäter olika konstrukt kan vara olämpliga att slå ihop i en metaanalys. Denna bedömning ska man göra med hänsyn till vilka slutsatser man vill kunna dra från resultatet. Som ett exempel, vill man veta hur stor effekt en behandling har på mängden frukt som en grupp personer äter är det rimligt att lägga ihop

mängden äpplen och päron, även om dessa är olika typer av frukt. Är man istället särskilt intresserad av mängden gröna frukter specifikt, kan det vara problematiskt att lägga ihop mängden päron och äpplen, eftersom äpplen ibland är röda.

Det samma gäller för konstrukt såsom depression. Studierna kan ha använt olika typer av skalor som alla är avsedda att mäta depression. Dock kanske skalorna skiljer sig åt så att vissa mätt suicidrisk och andra inte. Då kan man i projektgruppen exempelvis bedöma att frågan man vill utvärdera egentligen handlar om olika typer av depression (t.ex. lindrig depression, måttlig depression eller svår depression) och att man vill kunna dra separata slutsatser om de olika typerna. Man kan då behöva göra flera metaanalyser där man lägger ihop resultaten från endast de studier som använt samma skala eller mätt en viss typ av depression.

Oavsett om man väljer att lägga ihop studierna eller gruppera dem i undergrupper, baserat på vilken skala de använt, så kan man grafiskt presentera metaanalysen/metaanalyserna i ett grupperat format, genom att man sorterar studierna så att de som använt samma skala presenteras som en grupp. Om skalorna exempelvis är avsedda att mäta depression kan man gruppera de skalor som mätt suicidrisk i en grupp och skalor som inte mätt suicidrisk i en annan. Då blir det lättare att se hur homogena resultaten faktiskt är. Notera att grupperingen ska vara bestämd i förväg, innan man påbörjar analysen.

Steg 2: Skapa ett sammanslaget effektmått.

SMD – den vanligaste metoden

När man gör en metaanalys av studier som använt olika skalor kan det sammanslagna effektmåttet beräknas som ett standardiserat medelvärde (eng. standard mean difference, SMD). Att använda SMD som effektmått är en vanlig metod som möjliggör en sammanslagning av olika skalor. Detta görs genom att varje studies medelvärde delas med dess standardavvikelse. Därmed ändrar enheten för effektstorleken från antal steg på originalskalan till antal standardavvikelser.

En nackdel med att använda SMD är att det kan vara svårt att kvalitativt uppskatta effektens storlek när den beskrivs som ett visst antal standardavvikelser. Det kan därför vara värdefullt att komplettera ett resultat i SMD med ett eller flera alternativa effektmått för att underlätta tolkningen av effektens storlek.

En annan nackdel med att använda SMD, eller någon annan metod som använder sig av SMD som mellansteg, är att resultatet är beroende av standardavvikelser och därmed känsligt för sådana faktorer som påverkar dem. Tanken bakom SMD är att skillnaden i standardavvikelse beror på skillnaden mellan skalorna som används. Dock påverkar även faktorer som är orelaterade till denna skillnad mellan skalorna storleken på standardavvikelsen. Dessa faktorer kan därmed också påverka storleken på effekten, när den konverteras från originalskalan till SMD. Till dessa hör alla faktorer som kan ha påverkat spridningen i grupperna, exempelvis ett snävt eller liberalt urval av deltagare, eller antal personer i varje grupp. Om populationerna i de studier som ingår i en metaanalys skiljer sig åt kan alltså SMD vara ett missvisande effektmått. I sådana fall kan man med fördel också presentera resultatet på ett sätt som inte är beroende av standardavvikelsen.

Alternativa metoder som man kan komplettera ett SMD-resultat med

Ett resultat i SMD kan vara svårtolkat och därför kan det vara värdefullt att komplettera det med ett eller flera alternativa effektmått för att underlätta tolkningen av effektens storlek:

- **Konvertera från SMD till den mest använda skalan:**
  Ett resultat i SMD kan konverteras till den mest använda skalan, om det finns en sådan. Detta kan göras på metaanalysens resultat i SMD eller, med fördel, separat för varje enskild studie innan metaanalysen görs. I konverteringen använder man kontrollgruppens standardavvikelse vid baslinjemätningen, eller så poolar man alla studiers standardavvikelser för kontrollgruppen vid baslinjemätningen.
- **Konvertera från SMD till oddskvot:**
  Metaanalysens resultat i SMD kan också konverteras till en oddskvot för att underlätta tolkningen av effektens storlek. Oddskvoter är mycket väletablerade men konverteringen baseras på antaganden som inte alltid är uppfyllda (man antar normalfördelning med lika standardavvikelse mellan grupperna).

SMD som effektmått beräknas med hjälp av standardavvikelsen och därför är det känsligt för faktorer som påverkar standardavvikelsen. De effektmått som man beräknar utifrån SMD har även de samma nackdel. Därför kan det vara bra att presentera effektstorleken med ett effektmått som inte använder sig av standardavvikelser som ett komplement:

- **Omvandla en skala till en annan:**
  Ett sätt att omvandla en skala till en annan utan att använda sig av standardavvikelser är att omvandla den ena skalan så att den får lika många skalsteg som den andra. Detta görs för varje enskild studie och man kan därefter göra en metaanalys av dessa studier. Denna metod ställer dock höga krav på att skalornas max- och minimi-värden betyder samma sak då skalan i princip bara förlängs eller kortas ner till samma antal skalsteg som den skalan man vill konvertera till har. Om en studie exempelvis använt sig av 100 skalsteg och en annan av 10

skalsteg för att skatta samma symtom, såsom smärta, med samma värde för maximum och minimum (t.ex. "maximal smärta" och "ingen smärta alls") är det enda som egentligen skiljer skalorna åt hur fingradig skalan är. Det kan vara bra att omvandla den skala som är mest fingradig till den som är minst (dvs. från 100 skalsteg till 10 skalsteg), för att inte göra effektmåttet mer fingradigt än vad underliggande data är. Denna omvandling kan i detta fall vara lämplig och gör att man kan lägga samman studierna i samma metaanalys och ange effektstorleken i den skala som använts för att mäta effekten, vilket i sin tur underlättar tolkning av effektstorleken.

- **Beräkna Ratio of Means:**
  Ratio of Means (RoM) beräknar man genom att dividera medelvärdet i interventionsgruppen med medelvärdet i kontrollgruppen. Detta görs för varje enskild studie och sedan kan man göra en metaanalys med dessa uträknade värden. Den presenterade effektstorleken visar då hur många gånger bättre interventionsgruppen blivit jämfört med kontrollgruppen.
- **Beräkna en kvot mellan medelvärdesskillnaden och den minsta viktiga skillnaden:**
  För varje enskild studie kan man också beräkna en kvot med medelvärdesskillnaden i täljaren och den minsta betydelsefulla skillnaden (eng. minimal important difference, MID, minimal clinically important difference, MCID) i nämnaren. En metaanalys kan då göras med dessa värden och den presenterade effektstorleken visar då hur många gånger bättre interventionsgruppen blivit jämfört med kontrollgruppen, mätt i antal MID. Detta alternativ kräver dock väletablerade MID/MCID. Läsaren kan även behöva hjälp att tolka resultatet då en effekt som är mindre än MID lätt kan tolkas som att behandlingen inte har någon effekt. Det är lätt att glömma att hälften av individerna har en effekt större än medelvärdet och hälften mindre än medelvärdet (om man antar normalfördelning). Detta innebär att en stor andel kan få en effekt större än MID, trots att medelvärdet är mindre än MID. Det kan underlätta att sätta medelvärdet i relation till antalet individer som estimeras få en viss effekt.

Tabell 8.1 För och nackdelar med olika metoder för att skapa ett sammanslaget effektmått samt beräkningshjälp.

Metod	Fördelar och nackdelar (för en mer utförlig beskrivning, se [127])	Beräkningshjälp (för en mer utförlig beskrivning, se [126] [128])
SMD	**Fördel**: väletablerat **Nackdel:** svårtolkat och kan därför t.ex. inte användas för hälsoekonomiska beräkningar. Förutsätter att differensen i standardavvikelse mellan studierna beror på använd skala och inte skillnader i de studerade populationerna.	Detta görs lättast direkt i programmet RevMan.
Konvertera SMD till den mest använda skalan	**Fördel:** kan ses som närmare original-data. **Nackdel:** kan bli svårtolkat beroende på hur väletablerad skalan är. Konverteringen görs från SMD så de antagandena gäller även för denna metod.	Beräkna: SMD till skala
Skala om skalan till den mest använda skalan utan SMD	**Fördel:** kan ses som närmare original-data. Undviker de nackdelar som finns med att konvertera från SMD. **Nackdel:** kan bli svårtolkat beroende på hur väletablerad skalan är. Ställer höga krav på att skalans max- och minimi-värden betyder samma sak.	Beräkna: Skala om skalan till skala utan SMD
Konvertera SMD till en oddskvot	**Fördel:** mycket väletablerat och GRADE har riktlinjer till hjälp för tolkning av effektstorlek. **Nackdel:** baseras på antaganden som inte alltid möts (antar normalfördelning med lika standardavvikelse mellan grupperna). Konverteringen görs från SMD så de antagandena gäller även för denna metod.	Beräkna: SMD till OR
Ratio of Means (RoM)	**Fördel:** lätt att tolka, färre antaganden än andra alternativ. **Nackdel:** kräver att kontroll och interventionseffekten har samma tecken och är endast möjlig för post-testdata. Man måste även veta och kunna tolka kontrollgruppens medelvärde.	Beräkna: Skalor till RoM
Kvot mellan medelvärdesskillnaden och den minsta viktiga skillnaden (MID eller MCID).	**Fördel:** lätt att tolka, inte känsligt för populationsheterogenitet. **Nackdel:** måste ha tillgång till reliabel och väletablerad MID eller MCID för alla skalorna.	Beräkna: Skalor till MD i MiD

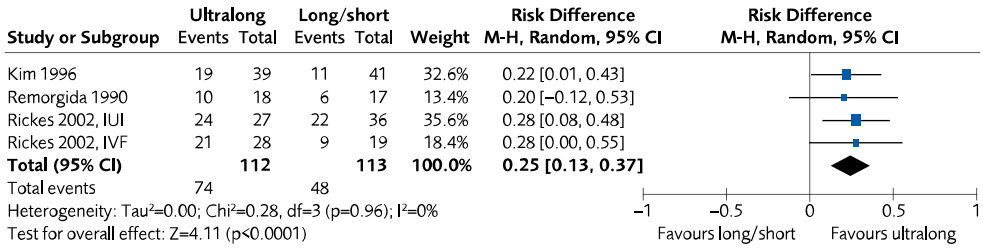
Faktaruta 8.4 Sammanfattande råd kring metaanalyser med kontinuerliga utfallsmått och tolkning av SMD.

- Om samma skala använts i alla studierna, använd medelvärdesskillnaden på skalan som effektmått.
- Om olika skalor använts i studierna, motivera varför de är tillräckligt lika (eller inte) för att slås samman i en metaanalys.
- Om du använder SMD som effektmått, komplettera det med ett annat effektmått för att underlätta tolkningen av effektens storlek.
- Ett sätt att underlätta tolkningen av SMD är att konvertera effektstorleken till en väletablerad skala, om det finns en sådan.
- Om det inte finns en etablerad skala så kan man underlätta tolkningen av ett resultat i SMD genom att konvertera det till en oddskvot.
- Om du använder SMD som effektmått, komplettera det med ett effektmått som inte är beroende av standardavvikelser. Ratio of means går oftast att beräkna och är inte beroende av standardavvikelser.
- Om det är möjligt, presentera både absoluta och relativa mått.
- Underlätta tolkning av effektens storlek genom att tydligt beskriva alla de skalor som använts i studierna.
- Förklara och motivera vad du gjort (t.ex. hur du gjort konverteringar och vilka data du valt för att göra dessa).
- Ge en ytterligare dimension till resultaten genom att beskriva dem i termer av antal individer som uppnått en viss effekt.

8.1.7 Forest plot

En metaanalys brukar presenteras som en så kallad forest plot (skogsdiagram). Diagrammet visar skattningar av effekt för de enskilda studierna, en sammanvägd effekt, konfidensintervall för såväl de enskilda effektskattningarna som för den sammanvägda effektskattningen, samt mått på heterogenitet. Figur 8.2 visar ett exempel på en forest plot med REM.

Figur 8.2 Exempel på en forest plot (skogsdiagram) [129]. Diagrammet visar andelen kvinnor med endometrios som blir gravida efter ultralång jämfört med kort GnRH-agonistbehandling, inför fertilitetsbehandling (IUI/IVF). Varje enskild studie listas efter försteförfattarens namn och publikationsår. Studiernas resultat redovisas som punktestimat (rektanglar) med tillhörande konfidensintervall (de horisontella linjerna). Storleken på rektangeln beror på den vikt som studien får i sammanvägningen. Det sammanvägda resultatet visas med en romb.



CI = Confidence interval; **M-H** = Mantel-Haenszel

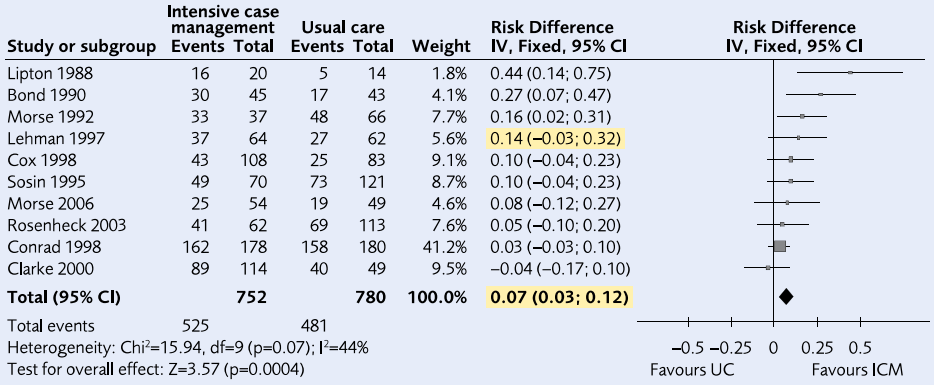
Fler fördjupande exempel finner du nedan.

Exempel på hur man arbetar med skogsdiagram

Figur 8.3 visar en forest plot, med resultaten av en intervention för hemlösa personer med psykisk funktionsnedsättning och mer eller mindre grava missbruksproblem [130-139].

Interventionen består av ett program kallat intensive case management (ICM) medan kontrollalternativet består av standardvård (UC för usual care). Effektmåttet är riskskillnad. Riskskillnaden anger här hur många procentenheter 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 till exempel tillfrisknande. Resultatet från varje enskild studie benämns enligt försteförfattaren, de horisontella linjerna visar konfidensintervallen och rektangeln i mitten visar vilken effektstorleken är.

Figur 8.3 Exempel på metaanalys (forest plot) – intensive case management (ICM) jämfört med standardvård (UC).



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

Diamanten (romboiden) längst ner visar den sammanvägda effekten samt konfidensintervallet för den sammanvägda effekten: en riskskillnad på 7 procentenheter och ett konfidensintervall från 3 till 12 procentenheter.

I kolumnen med rubriken Weight framgår vilken vikt respektive resultat har i sammanvägningen. Det "lättaste" resultatet (knappt 1,8 procent) kommer från en studie av Lipton och medarbetare [130] medan det resultat som väger tyngst har presenterats i en studie av Conrad och medarbetare (41,2 procent) [138]. Notera att ett resultat väger tyngre ju kortare konfidensintervallet är. Detta beror på att ju större standardfelet är, desto längre blir konfidensintervallet.

Figuren 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 8.3). 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 minskar att man missar en "sann" effekt på grund av att antalet ingående individer är för litet (risken för typ 2-fel eller b-fel minskar vid metaanalys eftersom den statistiska teststyrkan ökar).

Det finns emellertid några problem som gör att den sammanvägda effekten i Figur 8.3 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 på grund 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 till exempel populationens sammansättning, lokal kontext (sammanhang), interventionernas exakta innehåll, kontrollvillkoren, sättet att mäta effekterna, samt studiedesign. Detta problem brukar kallas klinisk heterogenitet [132] 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.

Även om alla resultat utom ett i Figur 8.3 uppvisar en positiv effekt är inte resultaten samstämmiga. Exempelvis varierar effektstorleken en hel del, från 44 procentenheter (Lipton 1988) till minus 4 procentenheter (Clarke 2000). Det går att kvantifiera denna bristande samstämmighet med olika mått på heterogenitet såsom I2 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 $\chi 2$ -test (Chi2-test) framgår att heterogeniteten är statistiskt signifikant i exemplet eftersom p=0,07, det vill säga <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 I2, 44 procent i fallet ovan. Annorlunda uttryckt, I2 utgör andelen av den totala variansen som förklaras av att det finns reella skillnader i effektstorlekar studierna emellan. Enligt en omtvistad tumregel sägs heterogeniteten vara låg om I2 ligger runt 0,25, måttlig om I2 ligger runt 0,50 och hög om I2 ligger runt 0,75, men att använda I2 som ett direkt mått på heterogenitet

kan vara svårtolkat och rekommenderas inte.

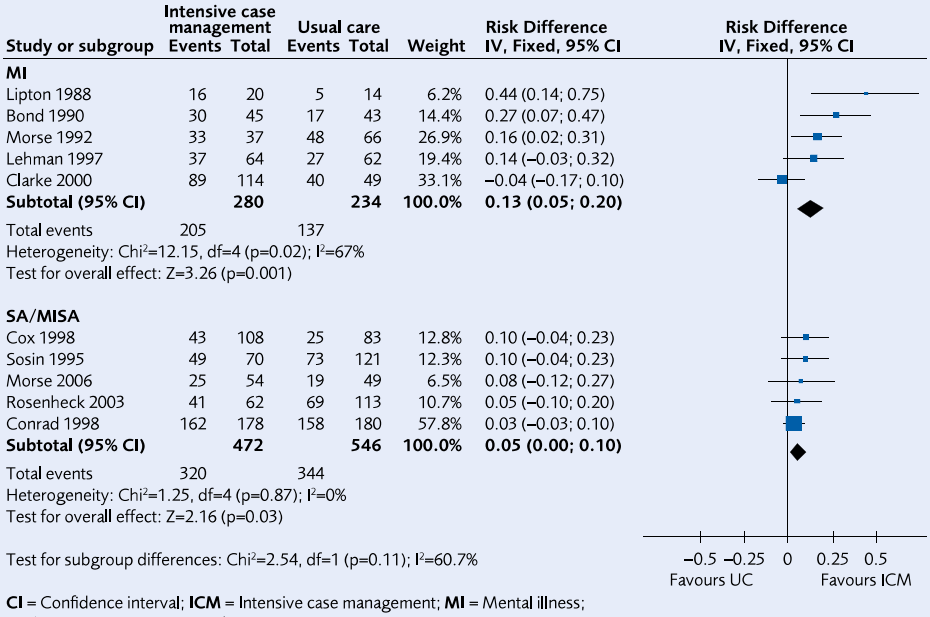
Anta att de olika resultaten bygger på studier som är mycket lika varandra avseende interventioner, kontrollvillkor, utvärderingsdesign och effektmått. Anta vidare att populationerna varierar från ett resultat till ett annat, men att positiva effekter trots detta uppvisar stor samstämmighet. Under sådana omständigheter tyder resultatet sammantaget på att interventionens skattade effektivitet är förhållandevis stabil oavsett subgrupper inom populationen (allt annat lika). I Figur 8.3 är resultaten emellertid inte samstämmiga vilket visar sig i den statistiska heterogeniteten.

Det kan finnas kliniska och metodologiska förklaringar till den bristande samstämmigheten. En möjlighet är att skilda patientgrupper reagerar olika på interventionen ICM. Den har i första hand utvecklats för personer med psykisk funktionsnedsättning, till exempel 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.

Subgrupper i en metaanalys

I Figur 8.4 har studierna och resultaten delats upp i två subgrupper utan total sammanvägning. Med denna gruppindelning framgår det att det inte finns någon heterogenitet inom SA/MISA-gruppen medan den till och med ökar inom MI-gruppen. Detta skulle kunna tyda på att ICM fungerar olika i de två grupperna, 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.

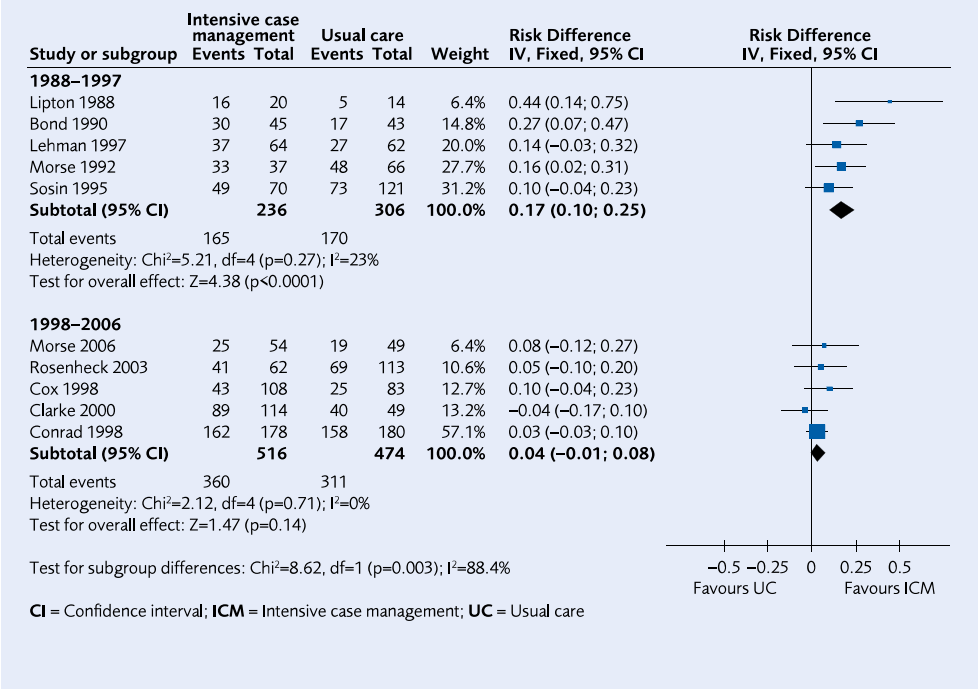
Figur 8.4 Subgrupper – psykisk funktionsnedsättning och drogmissbruk, intensive case management (ICM) jämfört med standardvård (UC).



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

Bakom heterogeniteten kan det finnas ett metodologiskt problem. Det kan uppstå när kontrollvillkoret utgörs av standardvård och den utvärderade interventionen består av ett antal mer eller mindre verksamma komponenter. Problemet orsakas av att komponenterna, som ingår i den nya och kanske mer effektiva interventionen, börjar spridas och integreras som delar av standardvården (en slags kontaminering). Vid kontaminering borde effekten av ICM i jämförelse med UC bli allt mindre över tid eftersom UC blir allt mer lik ICM över tid. I Figur 8.5 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.

Figur 8.5 Subgrupper – studier år 1988–1997 samt 1998–2006, intensive case management (ICM) jämfört med standardvård (UC).



8.1.8 Metaanalys av NRSI

Metaanalyser baseras huvudsakligen på resultat från randomiserade studier. Det går att göra metaanalyser som grundar sig på resultat från icke-randomiserade studier (NRSI), men det är ofta mer arbetskrävande. Grundprincipen är dock den samma. Man väger samman effekter där interventioner jämförs med kontrollvillkor. Ett problem är att sådana studier varierar mycket i vilken metodik som har använts. Variationen kan till exempel 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. Variationen kan också bero på antalet jämförelsegrupper och vid hur många tidpunkter mätningar görs.

8.1.9 Sammanvägning när underlaget består av både RCT och NRSI

Randomiserade och icke-randomiserade studier ska inte läggas in i samma metaanalys [47]. Om det finns randomiserade studier kommer de nästan alltid att ge ett tillförlitligare resultat. Om underlaget består av ett fåtal små randomiserade studier och flera stora icke-randomiserade studier kan man göra separata metaanalyser och undersöka om de visar samma eller avvikande resultat.

8.1.10 Val av programvara

SBU använder vanligen programmet RevMan från Cochrane Collaboration för att göra metaanalyser av interventionsstudier. Programmet är fritt tillgängligt via Cochrane. Undantagsvis krävs mera komplicerade beräkningar och då finns möjlighet att använda till exempel programmen Comprehensive Meta-analysis (CMA) eller R.

8.2 Metaanalys för diagnostisk tillförlitlighet

Studier om diagnostisk tillförlitlighet skiljer sig från interventionsstudier på flera sätt, vilket ställer andra krav på metoderna för metaanalys. Tre viktiga skillnader är:

- **Effektmåttet:** Sensitivitet och specificitet är beroende av varandra så att en ökad sensitivitet sker på bekostnad av en sänkt specificitet och vice versa. Metoden för metaanalysen måste kunna hantera två olika utfallsmått i en och samma analys.
- **Tröskelvärden:** Tröskelvärdet påverkar sensitiviteten och specificiteten. Ett lägre tröskelvärde kommer att medföra att sensitiviteten ökar. Omvänt leder ett högt tröskelvärde till att specificiteten ökar. Om studier har använt olika tröskelvärden måste metaanalysen kunna ta hänsyn till det.
- **Hög heterogenitet**: 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.

För att kunna utföra metaanalyser behövs därför metoder som tar hänsyn till både sensitivitet och specificitet, förhållandet mellan dem och heterogeniteten i testets tillförlitlighet (eng. test accuracy). Metoderna som används för metaanalys finns utförligare beskrivna i Cochrane Collaborations handbok för systematiska översikter [140].

8.2.1 Hierarkiska modeller

För att kunna ta hänsyn till den (oftast) negativa korrelationen mellan sensitivitet och specificitet, heterogeniteten och att studier använder olika tröskelvärden behövs multivariata metoder för metaanalys [140] [141] [142].

Två så kallade hierarkiska modeller har utvecklats för metaanalys av diagnostiska studier, den bivariata modellen och den hierarkiska sROC (HSROC) modellen. I benämningen ligger att modellerna består av två nivåer för att modellera data. På den första nivån behandlas variationen i sensitivitet/specificitet *inom* varje studie och på den andra nivån hanterar modellerna variationer *mellan* studierna.

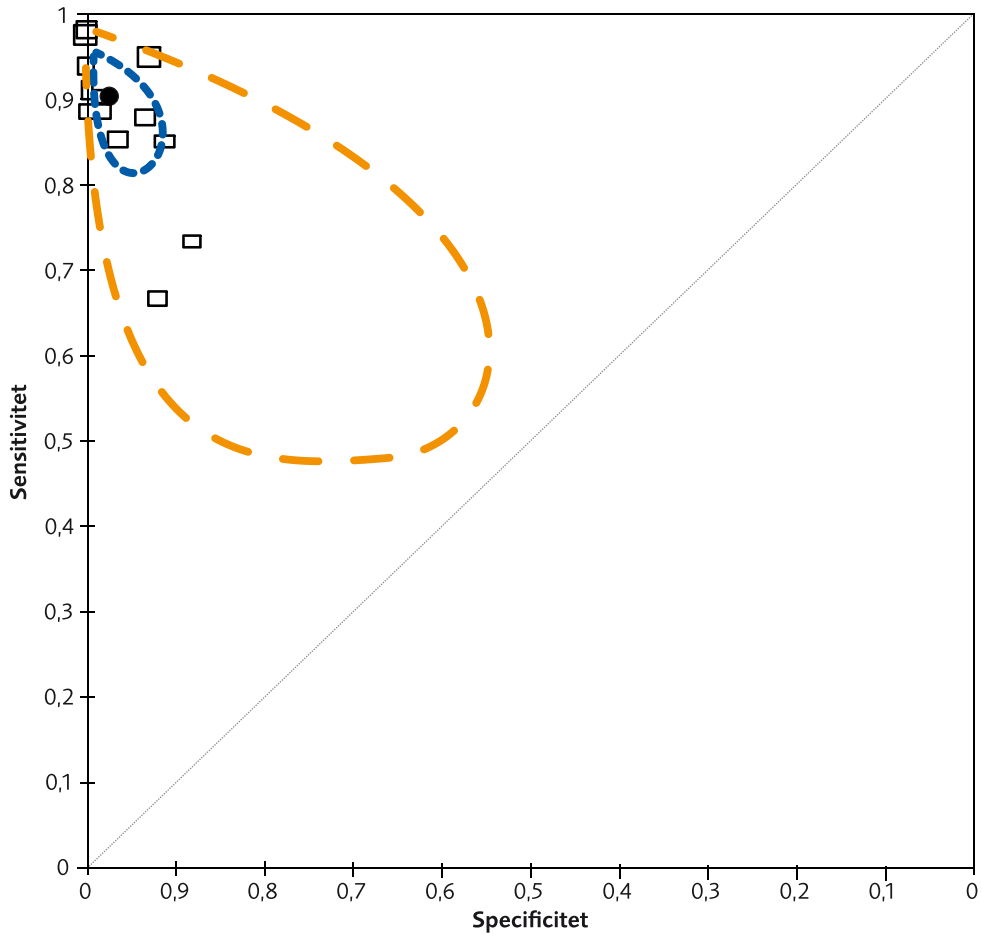
Parametrar som skattas med hjälp av båda modellerna kan därefter läggas in i RevMan. Resultatet blir antingen en sammanfattande punkt för sensitivitet och specificitet (bivariatmodellen) eller en sROC-kurva (HSROC).

8.2.1.1 Sammanfattande punkt

Metaanalys som ger en sammanfattande punkt (sammanfattande sensitivitet och specificitet), så kallad bivariat analys, används när resultaten bygger på samma tröskelvärde. Förutom punkten ger metaanalysen en 95-procentig konfidensregion och en 95-procentig prediktionsregion, se Figur 8.6. Konfidensregionen baseras på konfidensintervallet för den sammanfattande

punkten. Prediktionsregionen uppskattar området inom vilket vi skulle förvänta oss resultat från en framtida studie. Den är därför bredare än konfidensregionen. Konfidens- respektive prediktionsregionen är användbara för att illustrera osäkerheten i punktens värden och graden av heterogenitet.

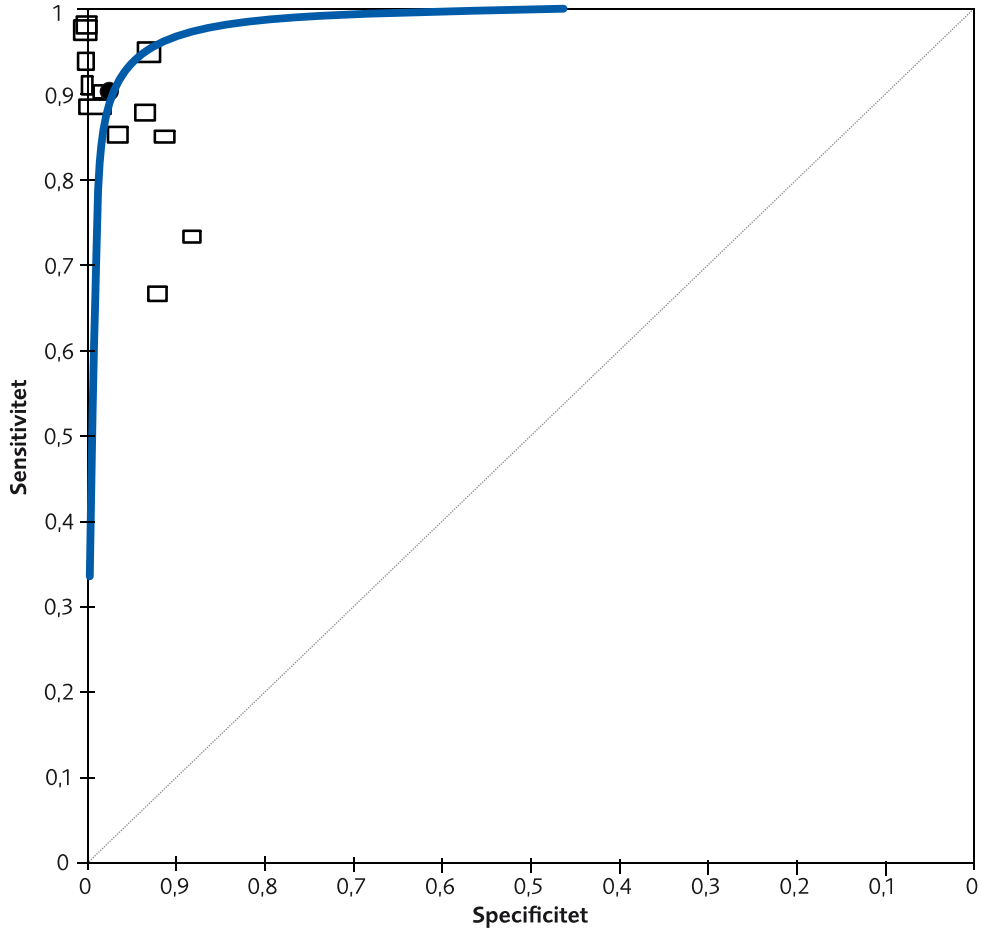
Figur 8.6 Exempel på bivariat analys. Den inre regionen (blå) är en 95 % konfidensregion medan den yttre regionen (orange) är en 95 % prediktionsregion.



8.2.1.2 HSROC-kurva

När resultaten i studierna baseras på olika tröskelvärden beräknas istället en summerande hierarkisk ROC-kurva (HSROC-kurva), se Figur 8.7. Linjen ska tolkas som att den beskriver hur den genomsnittliga sensitiviteten relaterar till den genomsnittliga specificiteten. Att punkterna är utspridda längs hela ROC-arean är att förvänta eftersom olika tröskelvärden användes. Däremot kan punkternas avstånd från kurvan ge en uppfattning om heterogeniteten – ju längre bort från kurvan punkterna ligger, desto mer heterogena är resultaten.

Figur 8.7 Exempel på HSROC-kurva (blå). Punkternas avstånd från kurvan ge en uppfattning om heterogeniteten – ju längre bort från kurvan punkterna ligger, desto mer heterogena är resultaten.



8.2.1.3 Sammanfattande punkt eller HSROC kurva?

Valet av hierarkisk modell beror på om den diagnostiska tillförlitligheten ska gälla ett visst tröskelvärde eller gälla över flera. Ibland kan det vara meningsfullt att beräkna både en sammanfattande punkt och en HSROC-kurva eftersom analyserna kan ge olika information och komplettera varandra.

Inkluderade studier kan rapportera resultat på olika sätt:

- **Alla studierna har använt liknande tröskelvärde:** Även när det är möjligt att definiera ett gemensamt tröskelvärde kommer det finnas större eller mindre variationer mellan studieresultaten. Variation kan uppstå på grund av skillnader i kalibrering av instrument, subjektiv tolkning av resultat samt skillnader i genomförande av testet. När alla studierna använt liknande tröskelvärde är det fördelaktigt att redovisa resultaten i form av sammanfattande punkt.
- **Varje studie rapporterar sensitivitet och specificitet för ett tröskelvärde men de har använt olika tröskelvärden:** I det här fallet är det meningslöst att presentera metaanalysresultat i form av sammanfattande punkt, utan ett lämpligt sätt blir en HSROC-kurva som beskriver hur sensitivitet och specificitet varierar med tröskelvärdet.

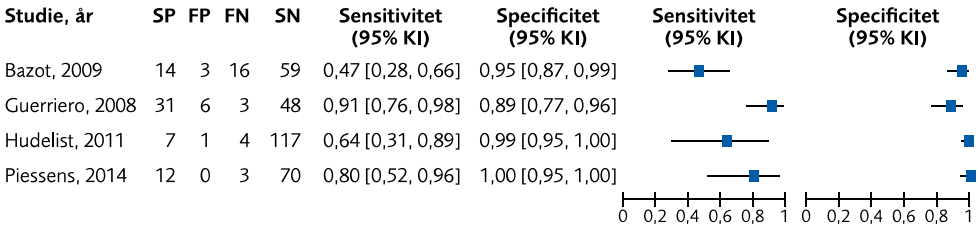
- **Några, eller alla studier rapporterar sensitivitet och specificitet för flera tröskelvärden:** Här kan man välja att räkna fram flera sammanfattande punkter, en för varje tröskelvärde. Man kan även välja att konstruera en HSROC-kurva över flera olika tröskelvärden – men observera att enbart ett tröskelvärde per studie får användas.

8.2.2 Heterogenitet

Heterogenitet i metaanalyser av diagnostiska studier är snarare regel än undantag. Testets sensitivitet och specificitet kan skilja sig åt mellan studier beroende på studiedesign, genomförande, sammansättning av deltagare, interventioner, indextestet, referenstestet och tröskelvärde. Därtill tillkommer heterogenitet som orsakas av slumpen och av bias som följd av brister i genomförandet av studierna.

En kopplad forest plot kan ge en snabb visuell överblick över heterogeniteten (se Figur 8.8). Ett annat sätt att undersöka heterogeniteten är att inkludera variabler som är karakteristiska till studierna (kovariater) i de hierarkiska modellerna. Kovariaterna kan exempelvis vara kön, ålder, blindning eller läkemedeldos, och genom att välja en kovariat åt gången kan vi studera dess effekt på effektestimat. Bivariatmodellen och HSROC-modellen skiljer sig i hur kovariaterna är inkluderade. För bivariatmodellen kan vi undersöka hur kovariaterna påverkar testets sensitivitet och specificitet. Med HSROC-modellen kan vi undersöka kovariaternas effekt på formen av sROC-kurvan och dess position i ROC-arean.

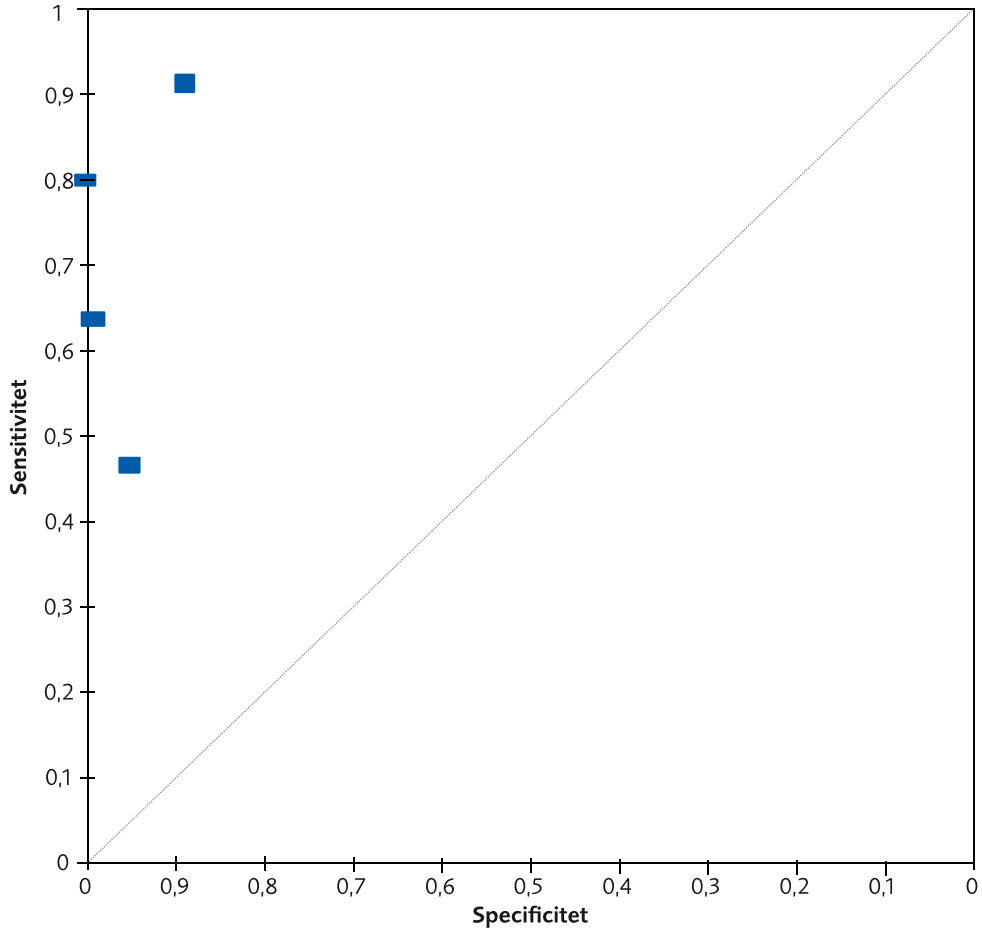
Figur 8.8 Kopplad forest plot kan användas för att få en uppfattning om studiernas heterogenitet.



8.2.3 Arbetsgång

- **Visuell inspektion av data:** Börja med att titta på sensitivitet och specificitet separat i exempelvis en kopplad forest plot (se Figur 8.8). Kopplad forest plot kan göras i RevMan genom att för varje studie lägga in antalet TP, FP, FN och TN. Om då studieresultaten uppvisar stor variation (t.ex. sensitiviteten i Figur 8.8) är det ingen idé att väga ihop studierna. Ett ytterligare sätt är att plotta studierna på en ROC-area (se Figur 8.9).

Figur 8.9 Studieresultaten inlagda i ett ROC-diagram.



- **Val av presentation av resultat:** Bestäm hur man ska hantera blandad rapportering av tröskelvärden som kan förekomma i studierna. Bör analysen begränsas till studier som delar ett gemensamt tröskelvärde (vilket möjliggör uppskattning av sammanfattande punkt) eller ska alla studier inkluderas, oavsett tröskelvärde (vilket möjliggör uppskattning av HSROC-kurva på bekostnad av tolkning av sammanfattande sensitivitet och specificitet)? Tänk på att båda modellerna kräver att det finns minst fyra studier att lägga in i metaanalysen. Om studierna visar 100 procent sensitivitet eller specificitet (s k nollceller) kommer modellerna inte heller att fungera optimalt. Om antalet studier är begränsat och/eller vid 100 procent sensitivitet och specificitet kan man försöka laborera med modellerna men det kräver en del programmering [143]. Eventuella subgrupper bör bestämmas innan man utför metaanalysen.
- **Metaanalys:** RevMan är inte anpassat för de hierarkiska modellerna utan delar av metaanalysen måste göras i ett annat statistikprogram, som till exempel SAS, Stata eller R som har speciella moduler för hierarkiska modeller. De enskilda studiernas resultat, uttryckt som sant och falskt positiva respektive sant och falskt negativa läggs in i statistikprogrammen. Resultaten från de multivariata analyserna importeras sedan till RevMan, som ger en grafisk redovisning av resultaten.
- **Undersökning av heterogeniteten:** Genom att göra en metaregression, det vill säga inkludera kovariater i modellen, kan man undersöka deras påverkan på den sammanfattande punkten eller HSROC kurvan. Observera

att sådan information inte alltid är tillgänglig i de inkluderade studierna. Kovariater kan läggas in i SAS och R men inte i Stata. Tänk på att vissa R-paket inte klarar av metaregression. För praktiska detaljer se Cochrane DTA handbok [140] eller Cochranes ”Software for meta-analysis of DTA studies”.

- **Tolkning av resultat:**
  - Sammanfattande punkt: Sensitivitet och specificitet presenteras ofta som proportioner eller procentsats. Vi kan göra resultaten mer begripliga genom att istället utrycka antalet TN/TP/FN/FP per 100 eller 1000 patienter. Exempelvis om sensitiviteten är 75 procent kan vi skriva att för varje 100 patienter med det sökta tillståndet, kommer vi att korrekt identifiera 75 patienter. Det är viktigt att specificera vilken grupp man syftar på (t.ex. per 100 patienter som testades eller 100 patienter med/utan tillståndet osv.). Vi kan även presentera resultat med hänsyn till tillståndets prevalens.
  - HSROC-kurva: Vi kan välja en punkt på kurvan vid en bestämd sensitivitet eller specificitet (t.ex. 95 %) och läsa av den motsvarande sensitiviteten/specificiteten (se Figur 8.7). Därefter kan man redovisa resultat på samma sätt som för den sammanfattande punkten. Det är viktigt att välja en punkt som är relevant för frågeställningen. Är det viktigare med hög sensitivitet eller specificitet?

Oavsett hur man presenterar resultaten är det viktigt att diskutera konsekvenserna av falskt negativa och falskt positiva resultat. Innebär det onödiga invasiva ingrepp vid falskt positiva resultat? Onödig oro? En försenad diagnos?

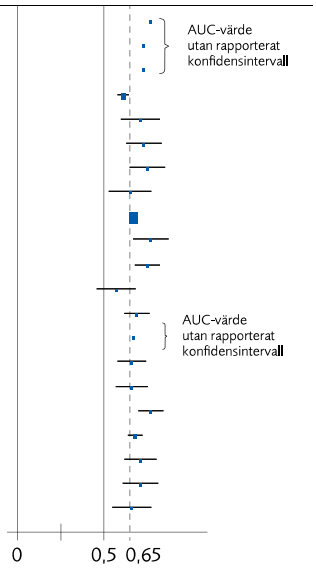
Med hjälp av en kalkylator i RevMan kan man få fram antalet TP/FN/FP/TN per 1000 testade patienter, beroende på tillståndets prevalens vid en viss sensitivitet och specificitet. De fält som är gröna fylls i för hand medan programmet räknar ut de övriga fälten. Om prevalensen är 0,2 och vi testar 1000 patienter, kommer vi korrekt identifiera 160 patienter med tillståndet, medan vi missar 40 patienter. Vi kommer vidare att korrekt friskförklara 760 patienter och ge falskt positivt svar till 40 patienter som inte har tillståndet.

8.2.4 Sammanvägning av AUC-värden

Ett mått som används i vissa studier är Area under the Curve (AUC). Att sammanväga AUC-värde är inte att rekommendera trots att flera modeller har föreslagits [144] [145]. AUC-värden som utgångspunkt för en analys bör endast göras när det inte är möjligt att beräkna sensitivitet och specificitet. En metod som prövats på SBU [19] är att lägga in de olika studiernas värden i RevMan och göra en visuell uppskattning av AUC, se exempel i Figur 8.10.

Figur 8.10 Hur man kan använda forest plot som stöd för att uppskatta AUC. Diagrammet skulle underlätta en bedömning av tillförlitligheten för att AUC var minst 0,65 [19].

Studie	Deltagare	AUC	Konfidensintervall 95 %
Åström et al, 2017	56	0,77	–
Catchpole et al, 2003	74	0,73	–
Catchpole et al, 2003	74	0,74	–
Chu et al, 2015	3 264	0,61	0,58; 0,64
Gammelgård et al, 2008	231	0,71	0,60; 0,82
Hilterman et al, 2014	105	0,73	0,63; 0,84
Hilterman et al, 2014	105	0,75	0,65; 0,85
Lodewijks et al, 2008b	117	0,65	0,53; 0,81
McGrath et al, 2018	4 401	0,67	0,65; 0,70
Meyers et al, 2008	121	0,77	0,67; 0,87
Olver et al, 2012	167	0,75	0,68; 0,82
Perrault et al, 2017	359	0,57	0,46; 0,68
Perrault et al, 2017	383	0,69	0,62; 0,76
Schmidt et al, 2005	104	0,67	–
Shepherd et al, 2014	213	0,66	0,58; 0,75
Shepherd et al, 2014	213	0,66	0,58; 0,75
Stockdale et al, 2014	147	0,77	0,70; 0,69
Vaswani et al, 2014	1 138	0,68	0,64; 0,73
Viljoen et al, 2017	156	0,72	0,63; 0,82
Viljoen et al, 2017	156	0,71	0,61; 0,81
Viljoen et al, 2019	108	0,66	0,55; 0,77



8.3 Narrativ sammanställning av kvantitativa data

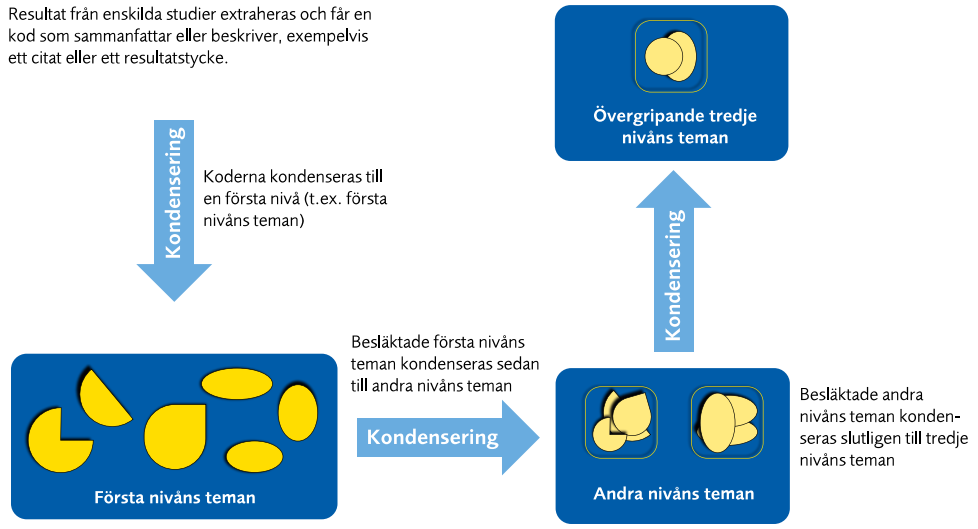
Om studierna som inkluderats är mycket heterogena, det vill säga att de skiljer sig åt med avseende på deltagare, intervention, kontrollåtgärd eller utfallsmått, på ett sådant sätt att det inte är rimligt att lägga ihop deras resultat, bör man inte göra en metaanalys. Istället får man tolka och sammanfatta resultatet med ord. En forest plot kan dock vara till hjälp när man tolkar resultaten. Den ska i så fall innehålla resultaten för varje enskild studie med samma utfallsmått, inklusive deras konfidensintervall. Diagrammet ger en visuell bild av materialet, vilket gör det mer överskådligt än om man redovisar effekterna enbart i separata figurer, eller i löptext. Någon sammanvägd effekt ska däremot inte räknas fram i detta fall, eftersom den ger en missvisande bild av att man trots allt har kunnat analysera studierna tillsammans.

8.4 Syntes av studier med kvalitativ ansats

Några år efter att kvantitativ metaanalys etablerats som en metod inom samhällsvetenskaplig forskning presenterades en motsvarande metod för kvalitativ syntes, metaetnografin [146]. Numera finns ett stort antal syntesmetoder beskrivna i litteraturen och i olika handböcker. Vissa syntesmetoder syftar till att beskriva ett fenomen utan att göra anspråk på att tolka resultaten. Andra metoder syftar till att tolka eller förklara, och ytterligare andra kan innehålla såväl beskrivande analys som en tolkning. I många metoder är målet att syntesen ska gå utöver (eng. go beyond) originalstudierna, det vill säga att syntesen leder till en helt ny tolkning som inte kan avläsas från de enskilda studierna [147]. Ingen syntesmetod ses för närvarande som ett givet förstahandsalternativ för HTA-rapporter men ofta kan metoderna meta-aggregering och tematisk syntes vara bra alternativ [24]. Gemensamt för alla metoderna är dock att de bygger på en stegvis kondensering, eller aggregering (se Figur 8.11).

Figur 8.11 Gemensamt för alla metoder som används för att syntetisera resultat från kvalitativa

studier är att de bygger på en stegvis process där resultat från enskilda studier kondenseras till övergripande teman.



EU har stött ett forskningsprojekt om kvalitativ syntes [89]. Projektet kom fram till att valet av syntesmetod påverkas av sju olika aspekter som sammanfattas i ramverket RETREAT (Review question, Epistemology, Time, Resources, Expertise, Audience and purpose, Type of data) [24] (se Figur 8.12). Mer information om RETREAT-kriterierna hittar du nedan.

Mer om RETREAT-kriterierna

Valet av metod för kvalitativ syntes påverkas av sju olika aspekter som sammanfattas i ramverket RETREAT (Review question, Epistemology, Time, Resources, Expertise, Audience and purpose, Type of data) [24].

1. Forskningsfrågan

Denna fråga är redan besvarad när syntesen påbörjas eftersom den handlar om hur forskningsfrågan ska formuleras (se avsnitt 3.4).

2. Epistemologi

Syntesmetoden beror på forskningsfrågan och på vilket synsätt översikten ska ha. Om syftet till exempel är att förstå ett socialt fenomen används troligen en annan metod än om syftet är att förstå effekter av en klinisk intervention. Med ett idealistiskt synsätt tenderar man att använda mer iterativa metoder för sökning och mindre fokus på granskning av studier. Ett realistiskt synsätt karakteriseras av en mer linjär och strukturerad process.

Inför valet av metod behöver projektgruppen överväga:

- I vilken utsträckning ska syntesen ta hänsyn till olika underliggande filosofier eller teorier i originalstudierna och hur ska skillnaderna hanteras? Meta-etnografi och grounded theory tar stor hänsyn till underliggande teorier.
- Vilket perspektiv har projektgruppen? Realist, idealist eller någonstans mellan? SBU:s projekt har generellt sett sådana frågor att ett realistiskt synsätt är det mest lämpade.

3. Tid

Den tid som projektet har till förfogande ska inte ensamt avgöra valet av metod men kan ändå spela en praktisk roll. I tid ingår dels när rapporten ska vara färdig, dels hur arbetskrävande metoden är. Faktorer som kan behöva beaktas är hur komplex metoden är, hur omfattande litteraturen är, hur många studier som inkluderas och hur rika och detaljerade data som de inkluderade studierna erbjuder.

Några syntesmetoder underlättar ett snabbt arbete. Meta-aggregation syftar till att presentera fynd från originalstudierna på ett tillförlitligt sätt utan att göra några tolkningar. Tematisk syntes ger möjlighet till att dels utveckla deskriptiva teman som ligger nära originalstudierna och dels, med mer tid till förfogande, att utveckla analytiska teman som är tolkande och genererar ny kunskap.

Frågor som behöver övervägas är:

- Syftar syntesen till att ta fram ny kunskap eller går det att använda existerande kunskapsresurser (kategorier, ramverk, modeller etc) för att snabba upp processen?
- Ska översikten bygga på en uttömmande täckning av alla studier som uppfyller inklusionskriterierna eller går det att snabba upp processen genom att använda ett strategiskt urval av litteratur?

4. Resurser
Här gäller frågan i första hand om det behövs eller finns tillgång till programvara som underlättar syntesen. SBU har hittills inte använt något specialiserat program för att strukturera kodning och kondensering utan genomfört syntesen med stöd av Word eller excel. Om syntesen ska göras som meta-aggregering tillhandahåller Joanna Briggs Institute programvaran SUMARI. Rad-för-rad kodning som ingår i en variant av tematisk analys kräver tillgång till program som NVivo eller Atlas Ti.

5. Expertis
Samtliga syntesmetoder kräver sakkunskap i stegen för en systematisk översikt samt ämneskunskap. Vissa metoder kräver djup förkunskap och erfarenhet av metoder som används i originalstudier.

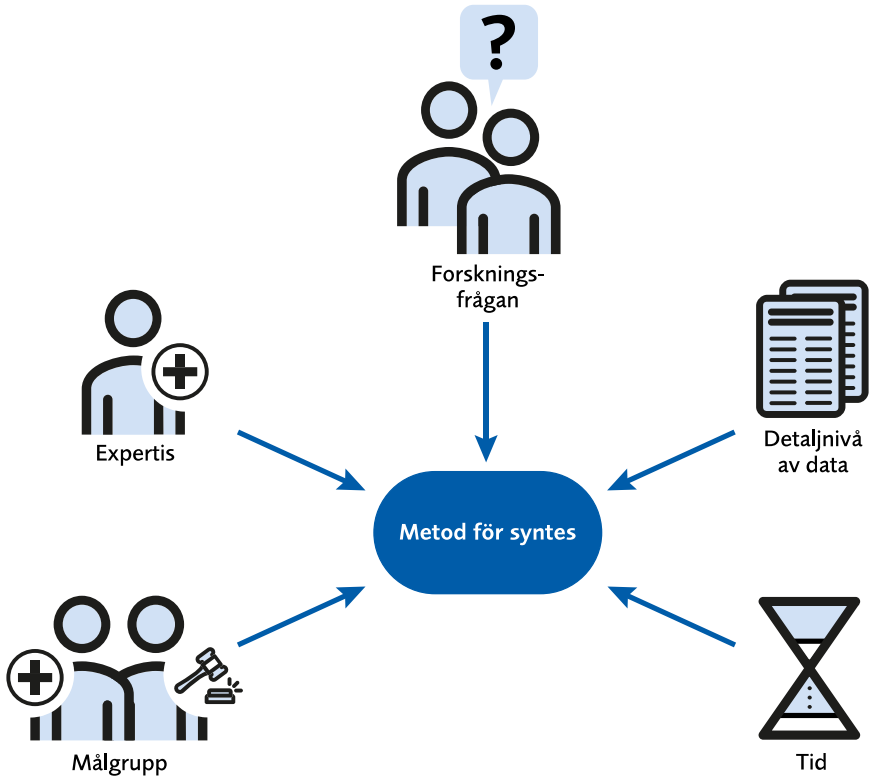
6. Målgrupp och syfte
Syfte och mottagare av rapporten spelar roll för vilka metoder som kan användas och hur fynden formuleras. Tematisk syntes, best fit framework-syntes och meta-aggregation utmynnar i resultat som formuleras på ett sätt som är mer direkt relevant för beslutsfattare än meta-studie och meta-etnografi som ger mera komplexa och konceptuella resultat.

7. Typ av data
Mängd och typ av data avgör om man ska välja en deskriptiv eller tolkande metod. Tolkande metoder fungerar bäst med ett mindre antal studier som detaljerat beskriver situation och kontext (eng. thickness) och/eller koncept (eng. richness), det vill säga i vilken utsträckning data är tillräckliga för att utveckla tolkande förklaringar eller utveckling av teorier. En syntes baserad på till exempel metoden metastudie kan genomföras med så få originalstudier som tre. Med större antal studier blir materialet så omfattande att det blir svårt att få en överblick men det finns inga definierade gränser för när studierna blir för många [24].

Det omvända gäller om studierna har "tunna" (eng. thin) data från till exempel enkäter med öppna frågor eller korta fallbeskrivningar. De kommer inte att vara tillräckliga för att tillåta tolkningar. Här krävs deskriptiva metoder som meta-aggregation, tematisk syntes, best fit framework- syntes och narrativ syntes som kan härbärgera större antal studier.

Sammanvägd bedömning
Om inte valet av metod blir tydligt efter att ha gått igenom de sju riskområdena rekommenderar INTEGRATE-HTA för närvarande att tematisk syntes används [89]. Metoden ger på sitt andra steg deskriptiva resultat men om data är tillräckligt rika går det att gå vidare till tolkande (analytisk) nivå.

Figur 8.12 Vad man ska tänka på inför val av metod för syntes av kvalitativa studier [24].



Avsnittet nedan beskriver kortfattat två syntesmetoder som är vanligt förekommande i HTA-rapporter: meta-aggregering och tematisk syntes. Även andra metoder kan vara tänkbara för ett SBU-projekt, under förutsättning att de sakkunniga har tillräcklig erfarenhet av metoden och ser den som bäst lämpad för att besvara forskningsfrågan. Systematisk innehållsanalys [148] [149] har till exempel använts i en SBU-rapport om rehabilitering för vuxna med traumatisk hjärnskada [149].

8.4.1 Meta-aggregering

Meta-aggregering är en textnära metod som lämpar sig väl när underlaget består av många studier med ”tunna” data. Programvaran SUMARI (The System for the Unified Management, Assessment and Review of Information) och dess verktyg QARI (Qualitative Assessment and Review Instrument) stödjer hela processen för meta-aggregering, inklusive granskning av studierna. Den granskningen är inte direkt användbar för bedömning av de syntetiserade fyndens tillförlitlighet med CERQual.

Metoden grundar sig i pragmatism och fenomenologi [150] [151]. Meta-aggregering är ingen tolkande analys av data från originalstudierna, istället koncentrerar sig metoden på ursprungsförfattarnas fynd i form av exempelvis kategorier och teman och sammanfattar gemensamma och motstridiga fynd över de inkluderade studierna så att de kan användas som grund för rekommendationer. Syftet är att balansera komplexiteten i ursprungsartiklarna med användbarheten av fynd för praktiker och beslutsfattare.

Studier som använder olika ansatser kan inkluderas i samma syntes. Fynd från studierna betraktas här som nivå 1 och därefter aggregeras dessa vidare till kategorier (nivå 2) och syntetiserade fynd (nivå 3). Ett praktiskt exempel på hur metoden tillämpas hittar du här [150].

Fynd på nivå 1 kan utgöras av ett tema, en kategori eller en metafor, som ska stödjas av ett illustrativt citat. Vid meta-aggregering ska man hålla sig så nära författarnas formulering av sina fynd som möjligt. Problemet är att en del fynd är allmänt hållna och därmed så breda att de kan bli oanvändbara för frågan. Det är då möjligt att ta tillvara mer specifika beskrivningar. Sådana fynd får då lägre trovärdighet, enligt metoden.

Den andra nivån undersöker likheter mellan fynd mellan de olika studierna och har ett tolkande inslag. Likartade fynd placeras i olika kategorier. Kategorierna bör skrivas som fullständiga meningar istället för enstaka ord som inte ger tillräcklig information. Kategorin ”klinikernas attityder var ett hinder för implementering av EBP” är till exempel mer användbar än ”attityder” enbart. Under respektive kategori finns en kort sammanfattning av de olika fynd som ingår. Enligt exemplet ovan skulle sammanfattningen kunna lyda ”kategorin avslöjar en brist på motivation eller vilja att arbeta evidensbaserat samt ett motstånd mot hela evidensrörelsen, vilket delvis hänger ihop med klinikernas personligheter och delvis till vissa discipliner inom sjukvården” [150].

Den tredje nivån är ett syntetiserat fynd som definieras som en övergripande beskrivning av en grupp fynd och som medger att man kan formulera rekommendationer baserat på dem. Fynden ska hjälpa till att överväga möjliga handlingsalternativ. I detta skede ingår också ett mått av tolkning. Ett syntetiserat fynd som omfattar kategorin om attityd kan bli: ”en brist på kompetens kommer att hindra implementering av EBP om gap i kunskap och färdigheter inte fylls och ansträngningar för att ändra kontraproduktiva attityder inte tas” [150]. I det beskrivna exemplet omvandlas fyndet till en rekommendation: ”integrera EBP i grundutbildningen” [150].

SBU har inte använt meta-aggregering för att syntetisera resultat från kvalitativa studier i någon rapport.

8.4.2 Tematisk syntes

Även den tematiska syntesen lämpar sig väl om underlaget består av stora mängder studier och ”tunn” data. Metoden används ofta för frågor om behov och för frågor om hur acceptabla och lämpliga interventioner är. Metoden, som utvecklades av Thomas och Harden [152], har ingen stark filosofisk komponent och studier inkluderas utan hänsyn till deras respektive teoretiska ansats. Studierna granskas med avseende på metodologisk stringens men samtliga relevanta studier tas med i syntesen. Därefter görs en sensitivitetsanalys för att undersöka om metodproblem slår igenom i resultaten.

Syntesen består av tre steg: 1) kodning av originalstudiernas fynd, 2)

konstruktion av deskriptiva teman och 3) utveckling av analytiska teman. De två första stegen är textnära (eng. data-driven) medan det tredje är teoridrivet. När forskningsfrågan handlar om till exempel behov kan frågan ses som ett teoretiskt ramverk.

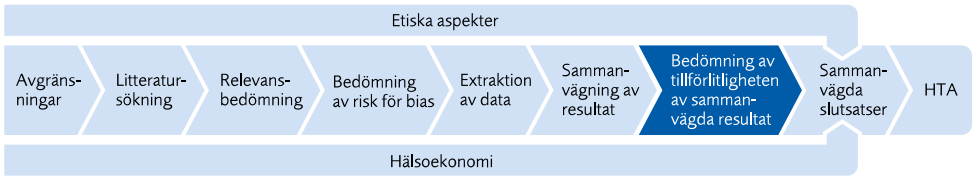
I det första steget sätts preliminära koder för varje rad av fynden (eng. line-by-line coding), för en studie i taget. Såväl citat som författarnas beskrivning av resultat kan utgöra meningsbärande enheter. Kodningen är induktiv, det vill säga inte bestämd i förväg. Koderna kan vara strukturerade hierarkiskt eller utan struktur. Denna del av arbetet underlättas om fynden läggs in i en programvara, till exempel NVivo. Vartefter kodningen fortsätter kommer man att bygga upp en bank med koder, även om nya koder kan tillkomma även vid kodning av de sista studierna.

I det andra steget söker man efter skillnader och likheter mellan koderna och grupperar dem i deskriptiva teman. I vissa fall kan det innebära att kodernas innehåll och namn förändras för att täcka hela innehållet. De deskriptiva temana sammanfattas i en text.

I det tredje steget utvecklas analytiska teman som ska ge ny kunskap, det vill säga ett tolkande steg. Teman utvecklas i en iterativ process som inkluderar länkar mellan deskriptiva teman och vilka följder de skulle kunna ha för deltagarna. De analytiska temana kan baseras på flera deskriptiva teman och varje deskriptivt tema kan förekomma i flera analytiska teman.

SBU har modifierat metoden och använt den i två rapporter. I den ena rapporten om myalgisk encefalomyelit och kroniskt trötthetssyndrom (ME/CFS) var slutsteget deskriptivt [153], medan man i den andra rapporten undersökte upplevelser om läkemedelsbehandling av kronisk smärta hos äldre, hos såväl de äldre som hos vårdpersonalen [154]. I denna rapport blev det sista steget därmed analytiskt där upplevelser beskrivna av de äldre triangulerades mot upplevelser beskrivna av vårdpersonalen.

9. GRADE –tillförlitlighet för sammanvägda resultat från kvantitativa studier



9.1 Introduktion

Det sista steget i processen är att bedöma hur tillförlitligt det sammanvägda resultatet är, det som tidigare kallades evidensgradering. SBU tillämpar GRADE (Grading of Recommendations Assessment Development and Evaluation) [155] som stöd för bedömningen av resultat från kvantitativa studier. SBU utgår från de principer som beskrivs i GRADE Handbook [156]. Mer information om GRADE finns på GRADE Working Groups webbplats.

Det sammanvägda resultatet kan uttryckas på flera sätt, ofta i form av ett punktestimat med sitt 95-procentiga konfidensintervall. Även tillförlitligheten för resultat från narrativ sammanställning kan bedömas med GRADE.

GRADE syftar till att på ett strukturerat och transparent sätt bedöma osäkerheter, risker, i det sammanvägda resultatet. En GRADE-bedömning görs därmed per utfallsmått. Till skillnad från äldre system är inte kvaliteten på ingående studier den enda utgångspunkten för en bedömning av om resultatet är tillförlitligt. Bristande samstämmighet mellan ingående studier och problem med överförbarhet är några andra faktorer som påverkar tillförlitligheten. GRADE kan ses som ett teoretiskt ramverk där resultatet granskas ur olika synvinklar (riskområden, eng. domains).

Med GRADE klassificeras tillförlitligheten som hög (⨁⨁⨁⨁), måttlig (⨁⨁⨁◯), låg (⨁⨁◯◯) eller mycket låg (⨁◯◯◯). Om det saknas studier som uppfyllt inklusionskriterierna klassificerar SBU det som GRADE (0). Beskrivningen av de olika nivåerna hittar du i Faktaruta 9.1.

Bedömningen inleds utifrån antagandet att resultatet har hög tillförlitlighet, det vill säga ⨁⨁⨁⨁. Det motsvarar att underlaget består av studier med optimal design för att besvara frågan, till exempel randomiserade studier om frågan gäller effekter av interventioner. Tidigare var utgångsbedömningen för NRSI att resultatet har låg tillförlitlighet (⨁⨁◯◯), på grund av risken för confounding. Eftersom confounding numera hanteras inom risk för bias är utgångsläget även för NRSI⨁⨁⨁⨁ [157].

Därefter bedöms risken för att resultatet påverkats av osäkerheter i de olika

riskområdena. Om osäkerheten som introduceras i ett riskområde är allvarlig så kommer tillförlitligheten att sänkas med ett steg. Anses osäkerheten vara mycket allvarlig så sänks tillförlitligheten med två steg. För NRSI tillkommer möjligheten att tillförlitligheten ökar, till exempel om effekterna är stora. Observera att en brist i underlaget ibland kan ge avtryck i flera riskområden. Det får då inte bli en ”dubbelbestraffning” så att det görs avdrag flera gånger för samma problem.

Faktaruta 9.1 Resultatens tillförlitlighet klassificeras i en av fyra nivåer enligt GRADE.

- Det sammanvägda resultatet har hög tillförlitlighet (⨁⨁⨁⨁)
- Det sammanvägda resultatet har måttlig tillförlitlighet (⨁⨁⨁◯)
- Det sammanvägda resultatet har låg tillförlitlighet (⨁⨁◯◯)
- Det sammanvägda resultatet har mycket låg tillförlitlighet (⨁◯◯◯)
  (Det går inte att bedöma om resultatet stämmer)

När det helt saknas studier som uppfyller inklusionskriterierna anges "studier saknas", utan gradering av tillförlitligheten.

Detta kapitel beskriver de olika riskområdena i GRADE liksom hur resultaten ska presenteras i en så kallad Summary of findings-tabell (SoF-tabell). Det kan inte nog betonas att GRADE är ett stöd för en strukturerad bedömning och att bedömningarna alltid kommer att ha subjektiva inslag. GRADE bidrar med att motiveringar och överväganden för bedömningarna ska framgå i SoF-tabellerna.

Till sist, GRADE är avsett för såväl resultat från systematiska översikter som för rekommendationer i riktlinjer baserade på systematiska översikter. GRADE skiljer mellan en systematisk översikt, som förutsätts vara oberoende av sammanhang (eng. context) och rekommendationer, som är beroende på sammanhanget. För SBU med uppdrag att göra utvärderingar för svenska förhållanden utan att göra rekommendationer innebär det att vi delvis följer GRADE för systematiska översikter och delvis för riktlinjer.

9.2 Riskområde 1: Risk för bias

Detta riskområde gäller inte risken för bias i enskilda studier, som redan är granskade med stöd av mallarna i Kapitel 6, utan hur stor risken är att det (sammanvägda) estimatet påverkas av brister i studierna [158]. Ett praktiskt hjälpmedel för att bedöma denna övergripande risk är en sammanställning av riskerna över samtliga inkluderades studier, det vill säga tabellen över risk för bias (se Figur 6.2).

En tumregel för bedömning är att inte göra ett enkelt genomsnitt av bedömningarna av respektive studie. Om det till exempel finns två studier som har många mycket allvarliga risker och två som har få och mindre allvarliga risker så ska man inte ge totalbedömningen ”allvarlig risk” och dra ner tillförlitligheten ett steg. Istället måste man noggrant överväga hur mycket varje studie bidrar till resultatet. Ett sätt är att utesluta studien ur metaanalysen och se hur mycket det

påverkar resultatet. Om studier med mycket allvarliga brister bidrar litet så påverkar de inte heller resultatet i avsevärd omfattning. I övervägandena ingår hur stora studierna är och antalet utfall (händelser). GRADE rekommenderar en försiktig hållning vad gäller att göra avdrag för risk för bias. Man ska ha en välgrundad uppfattning om att det finns en avsevärd risk för bias i de flesta studierna för att dra av.

Som regel inkluderar inte SBU studier med total hög risk för bias i sina analyser. Om man använder systematiska översikter genomförda av andra forskare där studier med hög risk för bias inkluderats kan det vara värt att överväga att ta bort dem från analysen om de förefaller störa resultatet. Nackdelen är att precisionen försämras eftersom antalet deltagare minskar.

9.3 Riskområde 2: Bristande samstämmighet

Med bristande samstämmighet, heterogenitet, avses att studierna visar olika resultat [159]. Om effekten varierar kraftigt mellan studier kan förklaringar ligga till exempel i att deltagarna haft olika svårighetsgrad av ett tillstånd, att interventionerna eller kontrollinterventionerna inte varit tillräckligt lika, att resultaten mätts vid olika tidpunkter eller att studierna haft olika risk för bias (högre risk kan förknippas med högre effektstorlekar [159]).

Om den bristande samstämmigheten inte kan förklaras minskar resultatets tillförlitlighet. GRADE skiljer mellan riktlinjer och resultat från en översikt i bedömningarna av samstämmighet. För systematiska översikter ställs hårdare krav på att studierna visar likartade effektstorlekar. För riktlinjer kan det vara tillräckligt att studierna samstämmigt visar till exempel negativ eller positiv effekt av en intervention, även om effektstorleken kan variera.

Bedömningen av samstämmighet beror följaktligen på om syftet med analysen är avgöra om det finns någon effekt över huvud taget eller att avgöra hur stor effekten är.

Några hållpunkter för när samstämmigheten brister på ett allvarligt sätt kan vara om:

- Resultaten varierar kraftigt mellan studierna, de "spretar". Läs mer i denna artikel av Guyatt och medarbetare [159].
- Om det statistiska testet för heterogenitet (som testar nollhypotesen att alla studier har samma underliggande storleksordning på effekt) visar ett lågt P-värde kan man dra av för heterogenitet. Om det inte har ett lågt P-värde bör man bestämma heterogenitet med andra metoder. Anledningen är att statistiska test för heterogenitet har låg teststyrka och ofta missar ofta skillnader som faktiskt finns.

För att undersöka orsaker bakom de skilda resultaten kan man ibland välja att genomföra stratifierade analyser på subgrupper. De ska vara definierade redan i projektplanen och funktionellt motiverade, till exempel ha en bakomliggande

biologisk förklaring.

Om subgruppsanalysen ingår i en redan publicerad systematisk översikt föreslår GRADE att analysen undersöks med en uppsättning kriterier. Mer om kriterierna hittar du nedan.

GRADE-kriterier för att bedöma tillförlitlighet i subgruppsanalys när metaanalysen genomförts av någon annan [159]

- Författarna hade definierat hypoteser om subgrupper och deras riktning på effekt i förväg.
- Rimlig mekanism för subgruppseffekt.
- Skillnader i effekt mellan olika subgrupper ses inom studier snarare än mellan studier.
- Statistisk analys antyder att slumpen är en osannolik förklaring.
- Skillnaderna i effekt i en subgrupp är synlig genom studierna och med olika utfallsmått.
- Subgruppsanalysen är en av få testade hypoteser, alternativt har hanterat multiplicitetsproblem.

9.4 Riskområde 3: Bristande precision

När vi tittar på precision bedömer vi konfidensintervallet för det sammanvägda resultatet.

GRADE fokuserar på konfidensintervallet för den absoluta effekten vid bedömning av osäkerheter i precisionen [160]. För SBU:s del kan det, beroende på vilken fråga som undersöks, i många fall vara relevant att fokusera på de relativa effekterna. För relativa effekter kan konfidensintervallet dock bli brett även när resultatet baseras på ett stort antal deltagare om antalet händelser i kontrollgruppen är lågt. Man kan då överväga att i sådana fall utgå från konfidensintervallet för den absoluta effekten som grund för bedömning av precision [160].

GRADE tillämpar olika kriterier för att bedöma brister i precision, beroende på om underlaget är en systematisk översikt eller en rekommendation i en riktlinje. För systematiska översikter bedöms enbart bredden och läget på konfidensintervallet och detta är den vanligaste utgångspunkten för SBU.

Ibland kan dock SBU välja att utgå från GRADE:s kriterier för riktlinjer för att bedöma precisionen. Här ska man ta ställning till om precisionen är tillräcklig för att stödja ett beslut. Man bedömer fortfarande bredden och läget på konfidensintervallen men ställer dem i relation till tröskelvärden för positiva respektive negativa effekter. Tröskelvärdena bestäms utifrån uppfattningar om värden och preferenser, till exempel av brukare. För en närmare beskrivning av hur tröskelvärden konstrueras och används i GRADE hänvisar vi till denna artikel av Guyatt och medarbetare [160].

Även om konfidensintervallen förefaller betryggande smala och resultatet robust

kan det finnas en underliggande brist i precision. Det uppstår när antalet händelser eller antalet deltagare är lågt. I små randomiserade studier kan det till exempel fortfarande finnas en prognostisk obalans mellan grupperna. GRADE föreslår i dessa fall att precisionen även bedöms med stöd av Optimal Information Size (OIS). OIS för en metaanalys överensstämmer med det minsta antal deltagare som skulle behövas i en studie för att få tillräcklig statistisk teststyrka, utifrån förväntad relativ riskminskning och antalet händelser i kontrollgruppen, för att kunna påvisa en viss önskad effekt. OIS kan uppskattas med hjälp av tillgängliga verktyg för att beräkna nödvändigt antal deltagare, som finns on-line. För att beräkna OIS behöver α och β specificeras liksom Δ, det vill säga önskvärd effekt. Ofta används $\alpha = 0{,}05$ vilket motsvarar 95 procent konfidensnivå och $\beta = 0{,}2$ vilket motsvarar en statistisk teststyrka på 80 procent. GRADE tillhandahåller också ett diagram som visar vilket antal deltagare som behövs för att uppnå olika nivåer av relativ riskminskning, se mer i denna artikel av Guyatt och medarbetare [160].

9.5 Riskområde 4: Bristande överförbarhet

Överförbarhet innebär att resultatet från studierna kommer att vara likartat för det sammanhang som forskningsfrågan avser [161]. Brister i överförbarheten kan relateras till skillnader i population, skillnader i intervention, skillnader i utfallsmått samt indirekta jämförelser. GRADE särskiljer mellan systematiska översikter och riktlinjer så att systematiska översikter enbart berörs av problem med surrogatmått och indirekta jämförelser. SBU har dock valt att ta hänsyn även till skillnader i population och intervention.

9.5.1 Population och intervention

I GRADE finns det sällan skäl att göra avdrag för skillnader i population om det gäller grupper, såsom patienter eller brukare. Det ska då finnas tunga argument för att de biologiska mekanismerna skiljer sig så mycket åt att effektens storlek påverkas. Undantagsvis kan underlaget för ett resultat baseras på helt andra populationer. Exempel är biverkningar som undersöks på råttor och apor eller penicillinresistens som kan mätas i provrörsmiljö. I dessa fall minskar överförbarheten och GRADE föreslår avdrag med två steg.

Överförbarheten kan påverkas av skillnader i miljö (eng. setting) och hur en interventionen implementeras. Studier där interventionen getts av forskare med god kontroll på dess genomförande ger till exempel sannolikt bättre effekter än när interventionen implementeras och genomförs utanför forskarens kontroll, något som kan motivera ett avdrag.

9.5.2 Utfallsmått

Det finns två viktiga aspekter på valet av utfallsmått. Den ena är användningen av så kallade surrogatmått. GRADE bygger på att utfallet mäts med mått som är viktiga för patienten eller brukaren (se även avsnitt 3.1.4 om val av utfallsmått). Problemet är att utfall som dödlighet och svår sjukdom inträffar mer sällan. Det

innebär att det krävs större studier med längre uppföljningstider och därför väljs ofta indirekta mått istället. Några exempel är att mäta effekter av blodtrycksbehandling som förändrat blodtryck istället för hjärtinfarkt eller död i hjärt-kärlhändelse, eller att mäta effekter av osteoporosbehandling som bentäthet istället för frakturer. För att bedöma hur överförbart ett indirekt mått är måste man ta hänsyn till bland annat verkningsmekanismer och naturalförlopp. I vissa fall kan överförbarheten betecknas som mycket bristande, det vill säga med två stegs avdrag. Ett exempel är behandling av personer med njursvikt och hyperfosfatemi med fosfatsänkande läkemedel. Ett för patienten viktigt mått kan vara hjärtinfarkt. Surrogatmåttet förkalkning av kranskärl kan motivera ett stegs avdrag medan mätningar av omsättningen av kalcium och fosfat kan motivera två stegs avdrag.

Den andra aspekten gäller uppföljningstider som avviker från forskningsfrågan. Effekter vid korttidsmätningar kan ha litet värde för bedömning av effekter på litet längre sikt. Många interventioner för att förebygga psykisk ohälsa hos barn har till exempel enbart uppföljningstider på några få månader trots att de syftar till att minska problem på flera års sikt [10].

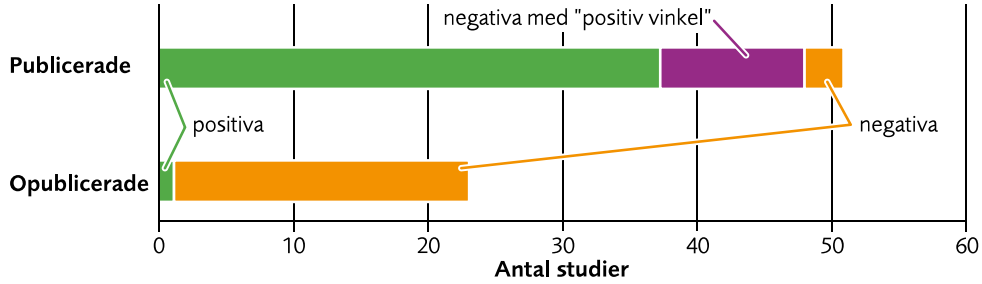
9.5.3 Indirekta jämförelser

Ytterligare en hörnsten i GRADE är att i första hand basera systematiska översikter på jämförelser mellan två interventioner som har förmodad effekt. Ofta saknas sådana direkta jämförelser (eng. head to head). Istället bygger underlaget på studier som till exempel jämför interventionerna var för sig mot placebo, eller mot ingen intervention. Enligt GRADE:s vägledning ska man då överväga att göra avdrag med minst ett steg för bristande överförbarhet [161]. Ett sätt att hantera problemet är att genomföra en nätverks-metaanalys.

9.6 Riskområde 5: Publikationsbias

Fenomenet publikationsbias har beskrivits i avsnitt 6.4.5.2 om ROBIS och är sannolikt mycket vanligt förekommande, oavsett om det gäller effekter av interventioner eller värdet av diagnostiska tester. Se Figur 9.1 för exempel.

Figur 9.1 Exempel på publikationsbias och hur man kan öka effektstorleken med 32 procent. [70]. Studien undersökte om effektstorleken för att minska symtom på egentlig depression med antidepressiva läkemedel påverkades av att resultat från opublicerade studier togs med i en metaanalys. Av de publicerade studierna visade huvuddelen av studierna att läkemedlen var effektiva. Ytterligare ett antal studier redovisade ingen signifikant skillnad på det primära utfallet men presenterade resultat för till exempel subgrupper ("positiv vinkel"). Av de opublicerade studierna såg endast två någon effekt av läkemedlen. Konsekvensen blev en överskattning av läkemedlens effekt.



Det är svårt att bedöma hur allvarlig risken för publikationsbias är. Det finns flera metoder som kan ge en fingervisning om att det saknas studier men det behövs indicier från mer än en metod för att göra avdrag. Till skillnad från övriga riskområden i GRADE kan man inte heller dra av med mer än ett steg. Riktlinjerna för GRADE rekommenderar att man överväger att göra avdrag med ett steg om underlaget enbart består av små studier [162]. Om de dessutom är sponsrade av företag eller om prövarna har någon annan form av intressekonflikter ökar risken för publikationsbias. Risken för publikationsbias kan man undersöka med hjälp av ett trattdiagram (se avsnitt 6.4.5.2, samt Figur 6.4).

En viktig informationskälla är om sakkunniga känner till att det finns studier som har presenterats på till exempel kongresser men som inte publicerats i vetenskaplig tidskrift. Ett bra komplement för interventioner är att undersöka om det finns några protokoll till studier registrerade i någon av forskningsdatabaserna, till exempel clinicaltrials.gov eller WHO:s databas ICTRP.

9.7 Att bedöma tillförlitlighet med GRADE när det bara finns en, eller ett fåtal studier

Tillförlitligheten i ett resultat ska bedömas med stöd av GRADE, även när det vetenskapliga underlaget är litet, det vill säga består av en enda studie eller ett fåtal små studier. GRADE tillämpas på samma sätt som när det finns ett mer omfattande underlag när det gäller överförbarhet och publikationsbias. Brister i samstämmighet är endast relevant att bedöma om det finns mer än en studie. Ett resultat som bygger på ett klent underlag blir dock mer känsligt för brister som leder till bias eller dålig precision, vilket beskrivs närmare nedan.

9.7.1 Ökar risken för bias när resultaten inte har upprepats av andra?

Risken för bias ökar om studien inte har upprepats av andra forskare eller forskargrupper. Ett undantag kan vara om underlaget består av en stor multicenterstudie där resultaten är samstämmiga mellan deltagande centra. Olika centra bör då ha bidragit i likartad utsträckning, det vill säga att det inte får vara så att ett enskilt, stort center fått en dominerande effekt på studiens resultat.

Farhågorna för att resultatet påverkats av bias minskar om det finns en vetenskaplig grund, till exempel prekliniska forskningsresultat, för den studerade interventionen, och inte bara en rimlig hypotes. Detta gäller särskilt om det finns en känd verkningsmekanism eller om interventionen bygger på en teoretiskt välunderbyggd och därmed allmänt vedertagen programteori.

På samma sätt minskar farhågorna när det finns vedertaget likartade interventioner inom samma område som har bekräftad effekt, till exempel läkemedel inom samma läkemedelsklass.

Slutligen minskar farhågorna om resultaten är likartade för olika utfallsmått, till exempel att samtliga visar en statistiskt signifikant effekt, eller om utfall med olika känslighet uppvisar samma trend. Om resultaten skiljer sig åt behöver inte det försvaga tillförlitligheten om det finns en bra förklaring, exempelvis om bortfallen för olika utfallsmått är olika stora.

9.7.2 Finns det risk för att förväntningar eller bristande forskningsetik påverkat resultatet?

När underlaget består av en enda studie eller av flera små studier där en enda forskare eller forskargrupp haft ett stort inflytande på genomförandet bör man vara extra uppmärksam på risken för att data har snedvridits. Studierna och de analyser som ingår kan ha vinklats för att bekräfta en viss hypotes och i värsta fall kan data vara fabricerade. Om projektgruppen bedömer att det finns en risk för felaktig rapportering kan det motivera ett extra avdrag i domänen Risk för bias, maximalt avdrag blir alltså –3.

Resultatet kan anses vara mer tillförlitligt när studien är gjord av forskare som inte själva utvecklat metoden eller interventionen som de studerar.

9.7.3 Är antalet observationer så litet att slumpen får en avgörande roll?

Det viktiga är inte hur många deltagare studien har utan hur många händelser som observerats. När det finns få händelser spelar slumpen en större roll. Det går dock inte att ge några generella råd om vad som är för få eller tillräckligt många observationer utan det får avgöras från fall till fall. Problem med få observationer hanteras inom domänen Precision.

Om den statistiska säkerheten i studien är övertygande med ett stort antal händelser så stärker det tillförlitligheten.

9.8 Faktorer som kan öka tillförlitligheten hos det sammanvägda resultatet

För kontrollerade studier utan randomisering kan det undantagsvis finnas skäl till att gradera upp tillförlitligheten ett eller två steg [163]. En förutsättning är att risken för bias inte får vara allvarlig.

Vägledningen till GRADE nämner tre faktorer som kan öka tillförlitligheten:

- Den sammanvägda effekten av en intervention är mycket stor.
- Det finns ett dos-responssamband.
- Det finns kända confounders som resulterar i en lägre effekt.

9.8.1 Stor effekt

Om det finns väl genomförda icke-randomiserade studier som visar en stor effekt är det troligt att effekten är verklig, och att det finns ett kausalt samband mellan interventionen och utfallet. Det är svårare att dra slutsatsen att effektstorleken är korrekt och inte en snedvriden överskattning. Enligt GRADE kan det finnas anledning att anta att det finns en effekt om:

- effekten (RR) ligger mellan 2 och 5 eller mellan 0,5 och 0,2 (ökar tillförlitligheten ett steg)

eller

- RR >5 eller RR <0,2 och det inte finns några allvarliga risker för bias eller brister i precision. Tilltron stärks ytterligare om effekten sätter in snabbt och att ett förlopp ändrar riktning (ökar tillförlitligheten två steg)

Anledningen till att det kan vara motiverat med en höjning är att modellstudier pekar på att confounding som följd av att randomisering inte har använts inte enbart kan förklara så stora effekter.

Om effekten beräknats som en OR kan samma kriterier användas som för RR om risken vid baslinjen är låg (typiskt lägre än 20 %). Om risken vid baslinjen är högre kan OR bli avsevärt större än RR. I sådana fall rekommenderar GRADE att man väljer en högre tröskel för effekt för att uppgradera.

9.8.2 Dos-responssamband

Ett orsakssamband stärks när ökad dos leder till ökad effekt.

9.8.3 Kvarvarande confounding minskar effekten

Även studier som hanterar kända confounders på ett exemplariskt sätt kommer att ha problem med kvarvarande confounding, det vill säga okända eller inte mätta prognostiska faktorer som har samband med utfallet. Under vissa omständigheter kan kvarvarande confounding leda till att effekten underskattas. Ett exempel är om bara sjukare personer får den experimentella interventionen men ändå förbättras mer än kontrollgruppen. Här är det troligt att effekten är underskattad. Flera exempel finns här [163].

På samma sätt kan man dra slutsatser om effekter när studier inte kan påvisa något samband. Här är ett exempel det tidiga fyndet att vaccination kan ge upphov till autism. Senare studier kunde dock inte bekräfta sambandet trots att

föräldrar till barn som fått diagnosen sedan den första artikeln publicerats var mer benägna att komma ihåg vaccinationen än föräldrar till barn utan autism. Att fyndet är negativt trots recall bias stärker tilltron.

9.9 Sammanställning i en SoF-tabell

Sammanvägda resultat för de olika måtten och deras tillförlitlighet ska redovisas i ett standardiserat format, en så kallad Summary of findings-tabell (SoF-tabell), se Tabell 9.1 [127] [164]. Syftet med tabellen är att underlätta för läsaren att förstå och tolka resultaten. Det måste framgå vad det är som bedöms med GRADE: är det ett punktestimat med konfidensintervall eller är det att det finns någon effekt över huvud taget? För läsaren kan det vara intressant att både relativa och absoluta effekter redovisas. Det kan också vara värdefullt att dela upp resultatet för deltagare med olika risker vid baslinjen. Relativa effekter är visserligen mer likartade oavsett risk, men uppgifter om absolut risk kan underlätta beslut i vård och socialtjänst.

Tabell 9.1 Exempel på en Summary of findings-tabell [154]. Tabellen sammanställer effekter av oxikodon på smärta vid diabetesneuropati hos äldre personer jämfört med placebo samt hur tillförlitliga resultaten är.

Utfallsmått	Antal individer respektive studier	Sammanvägt resultat	Tillförlitlighet i vetenskapligt underlag	Kommentarer
Oxikodon 10–160 mg jämfört med placebo				
Förändring på numerisk smärtskal (0–10)	n=497 2 RCT	Oxikodon minskar smärta med i genomsnitt 0,7 skalsteg (95 % KI, 0,29 till 1,12) mer än placebo	⊕⊕⊕◯ Måttlig tillförlitlighet för en *mycket liten*[1] effekt av oxikodon vad gäller smärta	Överförbarhet[2]: –1

[1] En effektskillnad vad gäller smärta med cirka 0,7 skalsteg på en skala 0–10 bedömer vi som mycket liten effekt.
[2] Bristande överförbarhet: studiedeltagarna var i genomsnitt cirka 60 år. Vår frågeställning berör individer 65 år och äldre.

KI = Konfidensintervall; **RCT** = Randomiserad kontrollerad studie; **RD** = Risk difference

För att inte överlasta sammanställningen rekommenderar GRADE att högst sju utfall ska ingå, eftersom större informationsmängder kan vara svåra att ta till sig. Om utfallet har undersökts i såväl randomiserade som icke-randomiserade studier och tillförlitligheten i resultaten är likvärdig bör båda resultaten tas med i tabellen. Om tillförlitligheten i resultat från två olika studietyper är olika ska det resultat som har högst tillförlitlighet läggas in i tabellen.

En viktig del av tabellen är att motiveringarna till varje GRADE-bedömning ska beskrivas i fotnoter. Det finns möjlighet att använda en programvara GRADE Pro som guidar igenom hur tabellen ska fyllas i.

Mer information om hur man lägger in uppgifter i SoF-tabellen för binära utfall finns här [164], och här kan du läsa mer om hur man lägger in kontinuerliga utfall [127].

Ett speciellt problem med att sammanställa resultat i tabellen uppstår när utfallsmåttet är kontinuerligt och beräknas som en standardiserad medelvärdesskillnad, uttryckt som SMD eller Cohen's d (se avsnitt 8.1). Den standardiserade medelvärdesskillnaden kan upplevas som svårtolkad. SMD kan dock översättas direkt till effektstorlek, uttryckt enligt tumreglerna för Cohen's d eller Hedge's g. För forskningsfält där de måtten är väl etablerade kan det vara en idé att presentera resultaten som Cohen's d (eller Hedge's g).

För att underlätta för läsaren att tolka SMD kan resultatet presenteras på fler än ett sätt [127]. Det är då viktigt att tydligt beskriva vilka formler eller data man använt sig av för att få fram dessa alternativ. Läs mer nedan.

Sammanställning i en SoF-tabell – olika alternativ för att komplettera ett SMD-resultat

För att underlätta för läsaren att tolka SMD kan resultatet presenteras på fler än ett sätt [127]. Det är då viktigt att tydligt beskriva vilka formler eller data man använt för att få fram dessa alternativ. Några alternativ är att komplettera ett resultat i SMD med att även:

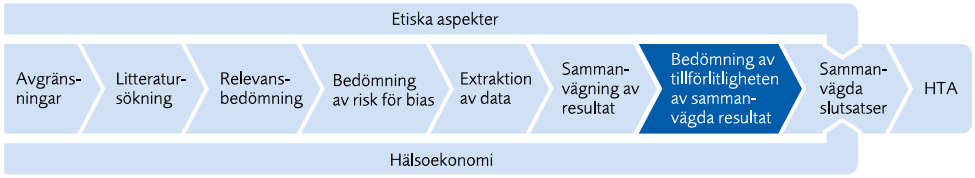
- Omvandla metaanalysens resultat i SMD till den mest använda skalan.
- Omvandla metaanalysens resultat i SMD till dikotomiserat format.
- Omvandla resultaten från varje enskild studie till den mest använda skalan och presentera resultatet från en metaanalys av dessa.
- Omvandla resultaten från varje enskild studie till en kvot av medelvärden (Ratio of means; RoM) där medelvärdet i interventionsgruppen delas med medelvärdet i kontrollgruppen. En metaanalys kan då göras med dessa värden och den presenterade effektstorleken visar då hur många gånger bättre interventionsgruppen blivit jämfört med kontrollgruppen.
- Omvandla resultaten från varje enskild studie till en kvot med medelvärdesskillnaden i täljaren och den minsta viktiga skillnaden (minimally important difference; MID) i nämnaren. En metaanalys kan då göras med dessa värden och den presenterade effektstorleken visar då hur många gånger bättre interventionsgruppen blivit jämfört med kontrollgruppen mätt i antal MID.

Samtliga alternativ som beskrivs i klickrutan ovan har allvarliga svagheter. En metod som använts i några SBU-projekt är att ta hjälp av ett etablerat värde på Minimal important difference, MID (se Kapitel 3). Man beräknar då andelarna för deltagarna i interventions-, respektive kontrollgrupperna i varje studie som förbättrats mer än MID, och väger sedan samman resultaten. Oavsett metod ska SMD redovisas i SoF-tabellen.

9.10 Diagnostisk tillförlitlighet

GRADE fokuserar på utfall som är viktiga för patienten eller klienten, det vill säga värdet av att en metod förbättrar hälsa eller minskar problem. GRADE anser därför att sensitivitet och specificitet är surrogatmått för det viktiga utfallet. Resultaten får därmed minskad överförbarhet. I de fall som SBU:s forskningsfråga gäller vilken diagnostisk tillförlitlighet en metod har, det vill säga när vårt eget primära utfallsmått är sensitivitet och specificitet, görs dock inget avdrag för brister i överförbarhet.

10. CERQual: Tillförlitlighet av resultat från en metasyntes



Tillförlitligheten för fynd från kvalitativa synteser bedöms med stöd av GRADE-CERQual [165]. Syftet är att på ett transparent sätt bedöma och beskriva hur stor tilltro som beslutsfattare och andra kan fästa på fynden. CERQual definierar tillförlitligheten som en bedömning av i vilken utsträckning fyndet är en rimlig representation av fenomenet. En alternativ formulering är i vilken utsträckning fyndet är ”substantiellt” skiljt från fenomenet. Med det menas att skillnaden är så stor att det påverkar beslutsfattandet.

Med fynd avses resultat från ett analytiskt arbete som, baserat på data från originalstudier, beskriver ett fenomen eller en aspekt av ett fenomen. CERQual är inspirerat av GRADE och har utvecklats i samarbete med GRADE Working Group. CERQual är avsett att fungera som ett strukturerat stöd för bedömningar och tolkningar som kommer att vara subjektiva. I publicerade studier har CERQual hittills tillämpats för deskriptiva fynd och inte för tolkande.

CERQual består av fyra riskområden: metodologiska begränsningar, relevans, koherens och tillräckliga data. På samma sätt som med GRADE utgår man från att fyndet är tillförlitligt och gör avdrag för brister som kan påverka tillförlitligheten. Tillförlitligheten klassificeras i fyra nivåer. Faktaruta 10.1 beskriver hur de olika nivåerna kan tolkas.

Faktaruta 10.1 Klassificering av tillförlitlighet enligt CERQual.

Nivå	Förklaring
Hög tillförlitlighet ⊕⊕⊕⊕	Det är mycket sannolikt att fyndet är en rimlig representation av fenomenet ifråga
Måttlig tillförlitlighet ⊕⊕⊕◯	Det är sannolikt att fyndet är en rimlig representation av fenomenet ifråga
Låg tillförlitlighet ⊕⊕◯◯	Det är möjligt att fyndet är en rimlig representation av fenomenet ifråga
Mycket låg tillförlitlighet ⊕◯◯◯	Det går inte att avgöra om fyndet är en rimlig representation av fenomenet ifråga

10.1 Riskområde 1: Metodologiska begränsningar

Med metodologiska begränsningar avses i vilken utsträckning design och

genomförande av studierna påverkar tillförlitligheten [114]. Bedömningen grundar sig på resultatet av granskningen av de individuella studier som är underlag för fyndet. Man måste ta hänsyn till hur mycket varje enskild studie bidrar, vilka brister som identifieras och hur de kan påverka fyndet. Här kan du läsa mer om metodologiska begränsningar [114].

Det går att göra en matris som illustrerar metodbrister över de olika studierna på samma sätt som för kvantitativa studier (se Figur 6.2).

10.2 Riskområde 2: Relevans

Med relevans avses i vilken utsträckning data från de underliggande studierna är tillämpliga för forskningsfrågan [116]. Detta motsvarar riskområdet Bristande överförbarhet i GRADE. Ofta stämmer studierna väl överens med satta inklusionskriterier men ibland måste man acceptera vissa avvikelser. Relevansen kan då bli indirekt, partiell eller osäker (Avsnitt 6.3.2). Här kan du läsa mera om relevans [116].

Bedömningen underlättas om relevansen i de enskilda studierna har bedömts i samband med granskningen av metodbrister.

10.3 Riskområde 3: Koherens

Kvalitativa fynd utvecklas genom att identifiera mönster i data över de studier som ingår. Med koherens avses att fyndet är väl underbyggt av data från studierna och ger en övertygande förklaring för mönstren [166]. Koherensen kan vara kontextuell, där studierna är likartade beträffande population, sammanhang mm eller konceptuell, där mönstren kan förklaras i relation till en ny eller existerande teori. Teorin kan vara internt utvecklad, det vill säga härröra från en eller flera studier i underlaget eller extern, det vill säga en etablerad teori. Ett tredje alternativ är att teorin utvecklas som del av syntesprocessen.

Fynd från synteser kan ses som transformationer av underliggande data till beskrivningar, tolkningar eller förklaringar av fenomenet. Beskrivningar är minst transformerade medan förklaringar är mest transformerade. Mellan dessa ytterligheter finns fynd som till exempel utforskar mönster av samband eller länka mönster i data till teoretiska koncept. Olika syntesmetoder ger fynd med olika grad av transformation. Meta-aggregation ger mer deskriptiva fynd medan till exempel metaetnografi ger mer förklarande fynd. Risken för bristande koherens ökar ju mer förklarande fynden är.

Deskriptiva fynd ger en sammanfattning av underliggande mönster av data i studierna. Om mönstren är komplexa eller varierande beror koherensen på hur väl komplexitet och variation beskrivs i fyndet. Det innebär att ett fynd kan behöva beskrivas detaljerat. Koherensen försämras om fyndet bara beskriver de mest dominanta mönstren och inte täcker oklara eller avvikande data. Ett exempel är fyndet ”kvinnor känner sig bekväma med att genomföra en medicinsk abort hemma” som är en alltför förenklad bild av fyndet ”kvinnors

erfarenheter av att genomföra en medicinsk abort i hemmet varierade. Några kände sig överväldigade, andra kände sig komfortabla och ”empowered” och ytterligare några uppgav att det var precis som vilken annan mindre procedur som helst” [166].

Koherensen i mer förklarande fynd minskar om det finns data i underliggande studier som utmanar tolkningen eller förklaringen eller om det finns möjliga alternativa tolkningar eller förklaringar.

Bedömning av koherens i en egen metasyntes ger en möjlighet till både reflexivitet och att överväga om det kan finnas andra sätt att syntetisera fynden som bättre kan fånga underliggande data. Man bör aktivt leta efter data som komplicerar eller utmanar fynden och försöka förklara dessa variationer eller undantag. Om man inte kan komma fram till någon övertygande förklaring till dem minskar tilltron till att fyndet representerar fenomenet. Det kan finnas flera orsaker till att det är svårt att förklara undantag, såsom att data kan vara för magert, teorin kan ha brister eller att urvalet av studier till översikten kan vara för begränsat.

Man ska undvika att släta över eller bortse från motstridiga fynd. Det kan vara frestande att till exempel formulera fyndet på ett vagare sätt för att öka koherensen men hela syftet med bedömningen är att klarlägga graden av osäkerheter i fyndet. Granskningsmallens fråga om koherens är ett stöd vid bedömningen.

Här kan du läsa mer om koherensbedömningen beskriven ovan [166].

10.4 Riskområde 4: Tillräckliga data

Bedömningen av tillräckliga data för de enskilda studierna i underlaget görs i Fråga 7 i granskningsmallen. Riskområdet handlar dels om hur rikt data är, dels kvantiteten data [167]. Rika data ger tillräckligt med detaljer för att man ska förstå fenomenet, men mängden data är också viktig. Om underlaget består av ett fåtal studier eller ett fåtal observationer minskar tilltron till att fyndet återspeglar fenomenet. Vi vet inte om studier som genomförs i andra miljöer eller med andra grupper skulle ge samma bild.

Det finns inga regler som avgör när data är tillräckligt rikt eller tillräckligt omfattande utan det blir en bedömning som görs från översikt till översikt. CERQual föreslår att begreppet mättnad kan vara användbart i vissa fall eller att man överväger i vilket utsträckning ytterligare data skulle påverka fyndet. För övrigt kan ett mindre antal konceptuellt rika studier bidra mer till ett fynd än ett större antal studier med magra, deskriptiva data. Läs gärna mer här kring bedömning av tillräckliga data [167].

10.5 Sammanvägd bedömning

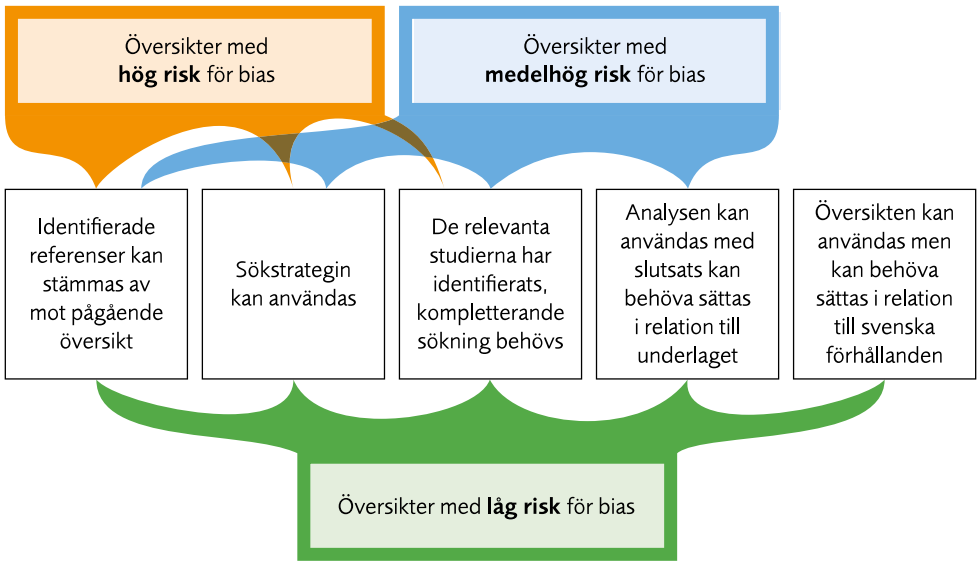
I likhet med GRADE sammanställs fynd, antal underliggande studier och

deltagare samt en sammanvägd bedömning av tilltron till fyndet i en SoF-tabell (se Kapitel 9). Motiven till avdrag ska framgå i anslutning till tabellen, till exempel i form av fotnoter.

11. Användning av redan publicerade systematiska översikter

Under de senaste 20 åren har antalet publicerade systematiska översikter och metaanalyser ökat kraftigt. Att återanvända publicerade systematiska översikter kan därför vara ett kostnadseffektivt arbetssätt som ökar hälso- och sjukvårdens samt socialtjänstens tillgång till evidensbaserad kunskap. Systematiska översikter från andra aktörer kan användas antingen helt eller delvis (Figur 11.1).

Figur 11.1 Möjliga användningsområden av andra aktörers systematiska översikter. Möjligheten att använda översikter med hög risk för bias varierar beroende på vilket delsteg i granskningen de uppfyller, alla går dock att använda för att stämma av identifierade referenser.

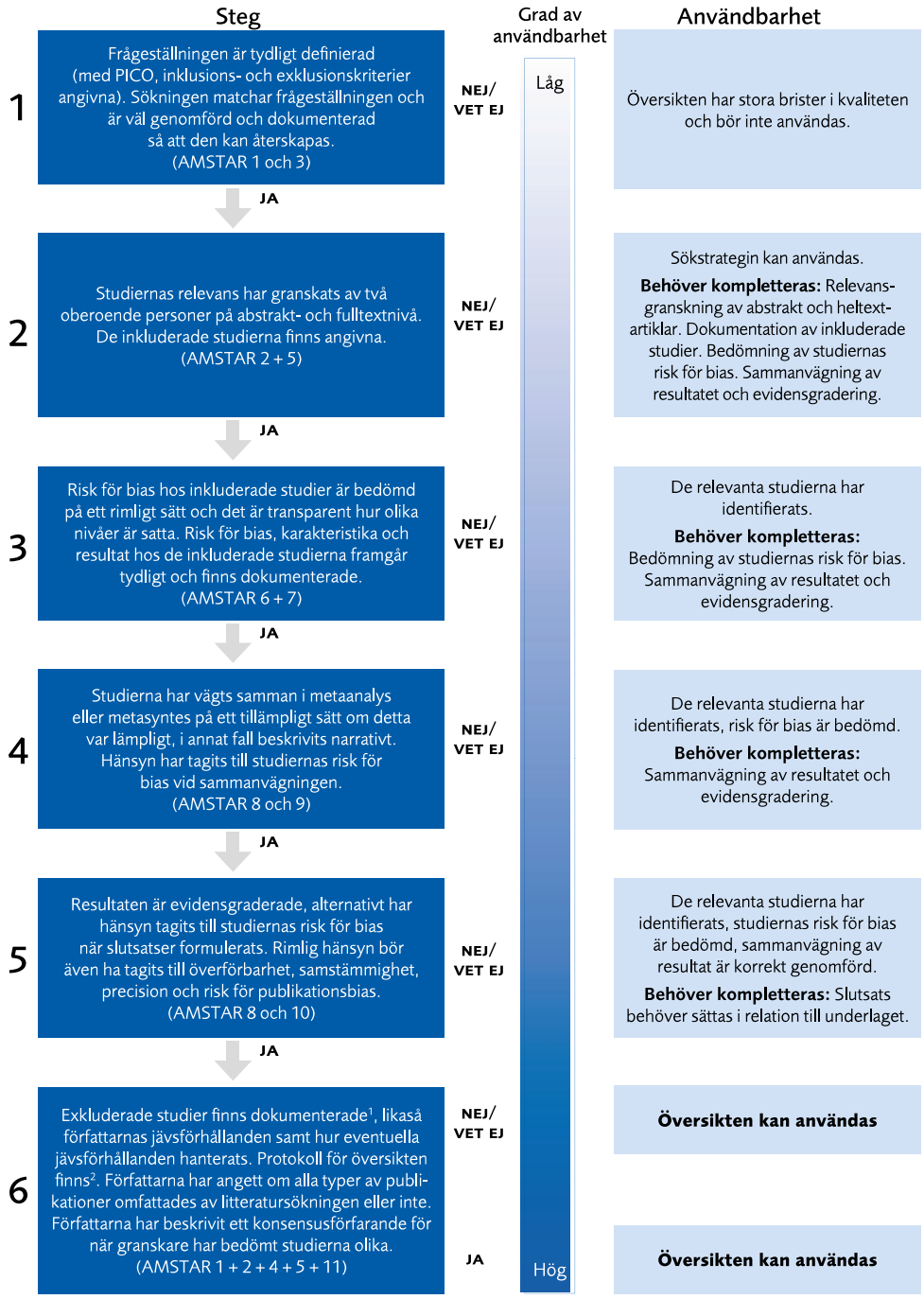


11.1 Systematiska översikter som enda underlag för en rapport

Förutsättningen för att kunna använda en redan publicerad systematisk översikt är att den uppfyller projektets inklusionskriterier. I ett första steg bedöms graden av användbarhet med hjälp av formuläret SNABBSTAR (se Figur 11.2), som bygger på frågorna i granskningsmallen AMSTAR [168] [169]. Syftet är att med minsta möjliga arbetsinsats avgöra vilka översikter som inte kan användas alls, vilka som kan vara grund för fortsatt arbete och vilka som redan är helt användbara. De översikter som bedöms som helt användbara bör granskas ytterligare med hjälp av ROBIS (se avsnitt 6.4) innan man inkluderar dem i sin översikt. Om man däremot enbart vill använda sig av sökstrategin eller inkluderade studier från en befintlig systematisk översikt behövs ingen ytterligare granskning. Man kan även använda sig av resultaten i en översikt men göra en ny bedömning av tillförlitligheten med hjälp av GRADE. För att användas ska översikten som regel inte ha mer än medelhög risk för bias, det vill säga att risken för bias i de ingående studierna ska ha granskats och resultaten ska vara syntetiserade på ett lämpligt sätt för att resultaten ska kunna användas.

SBU baserar i huvudsak sina metaanalyser på studier med låg eller måttlig risk för bias. Andra översikter kan ha valt att lägga in samtliga studier som är relevanta, med syfte att få så många studiedeltagare i analysen som möjligt. Några, till exempel Cochrane Collaboration, redovisar numera ofta även en metaanalys på enbart studier med låg risk för bias, det vill säga ett strängare urval än vad SBU tillämpar. Det är viktigt att avgöra från fall till fall vilka resultat som är användbara. En möjlighet är att acceptera metaanalyser som inkluderar studier oavsett risk för bias men ta hänsyn till det vid bedömning av tillförlitligheten till översiktens resultat med GRADE (se Kapitel 9).

Figur 11.2 Formuläret SNABBSTAR för översiktlig granskning av systematiska översikter, som bygger på frågorna i granskningsmallen AMSTAR [168] [169], i vilken graden av en systematisk översikts användbarhet bedöms utifrån frågorna i de blå rutorna.



[1] Det är viktigt att de exkluderade studierna finns angivna i anslutning till den systematiska översikten, eller i alla fall sammanfattning av skäl till exkludering. Det förekommer dock att dessa saknas beroende på begränsningar i utrymme hos vissa tidskrifter. SBU anser därför i dagsläget att en systematisk översikt kan bedömas ha medelhög användbarhet även utan att en lista på exkluderade studier finns tillgänglig.

[2] Det är viktigt att den systematiska översikten föregås av ett protokoll som stämmer överens med det som rapporteras i översikten. För de systematiska översikter som görs idag är det en naturlig del i processen, men för lite äldre översikter kan referens till protokollet eller själva protokollet vara svåra att finna. SBU anser därför i dagsläget att en systematisk översikt kan bedömas ha medelhög användbarhet även utan protokoll.

11.2 Systematiska översikter i utvärderingsprojekt

För utvärderingsprojekt, inklusive underlag till Nationella Riktlinjer och liknande, kan också systematiska översikter användas, helt eller delvis. Förutsättningen är att översikterna har samma PICO (eller motsvarande) och en låg risk för bias. Översikterna granskas med stöd av ROBIS (se avsnitt 6.4) eftersom resultatens tillförlitlighet kommer att bedömas med GRADE.

Följande principer ska tillämpas:

1. Projektledaren gör en preliminär bedömning av litteratursökningen. Översikter med uppenbara brister i sökstrategin exkluderas. Hit hör att endast en databas har använts eller att sökningen har baserats på ett antal beskrivna sökord, det vill säga det saknas en fullständig sökdokumentation.
2. Därefter bedömer två personer i projektgruppen översikten utifrån övriga riskområden i ROBIS. Om översikten bedöms ha låg risk för bias kontrollerar informationsspecialisterna att sökstrategin är tillräcklig.

Även systematiska översikter där resultatet har måttlig risk för bias kan användas om svagheten ligger i att författarna inte har bedömt risken för bias i enskilda studier. Den ökade osäkerheten hanteras i GRADE, med avdrag för risk för bias (se Kapitel 9). Förutsättningen är givetvis att litteratursökningen är godkänd.

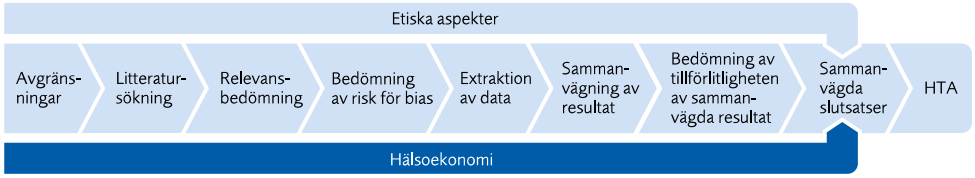
Flera SBU-rapporter, till exempel [170] [171] [172] har utgått helt eller delvis från ”äldre” systematiska översikter. Projektets litteratursökning tar vid där den systematiska översikten slutade.

Det kan också vara möjligt att använda delar av den systematiska översikten, beroende på vilka brister som finns. Ett exempel är SBU:s rapport om att förebygga missbruk hos barn och unga [173]. Här fanns flera översikter från Cochrane Collaboration med godkänd sökstrategi och tillfredsställande process för att gallra litteratur. Problemet var att rapporterna accepterat studier med kortare uppföljningstid än SBU:s PICO. Med ledning av information i tabellerna valdes de studier som hade tillräcklig uppföljningstid ut, risken för bias bedömdes och studier med låg och måttlig risk för bias fördes in i egna metaanalyser. I och med att det ursprungliga antalet abstrakts var mycket stort sparades mycket tid genom att använda Cochrane-rapporterna.

11.2.1 Flera översikter finns

För väl beforskade områden kan det hända att det finns flera systematiska översikter som har låg risk för bias. SBU har tidigare i utvärderingsprojekt utgått från de vägledande principer som Agency for Healthcare Research and Quality (AHRQ) i USA tagit fram [174] [175]. De kan sammanfattas som att den bästa (mest relevanta och med minst risk för systematisk bias) och senast publicerade översikten ska användas. Ett alternativt sätt, enligt AHRQ, är att enligt förbestämda kriterier för aktualitet redovisa de översikter vilka bedöms relevanta och med låg risk för bias. Den metoden kräver dock att översikterna är samstämmiga. Om de visar motsägande resultat kan det vara en tydlig signal att det behövs en ny oberoende systematisk utvärdering.

12. Hälsoekonomiska utvärderingar



12.1 Inledning

I SBU:s uppdrag ingår att utvärdera metoder ur ett ekonomiskt perspektiv. I kombination med att efterfrågan på sjukvård och omsorg hos befolkningen är hög och dessutom ökar [176][177] uppstår ett gap mellan vad samhället kan erbjuda och vad som efterfrågas. Därför behöver man göra prioriteringar mellan olika behandlingar, eller diagnostiska metoder, som resurserna ska läggas på. Hälsoekonomiska utvärderingar är ett stöd för beslutsfattare att avgöra huruvida en metod ger så pass mycket hälsa att det står i proportion till vad den kostar.

Hälsoekonomiska metoder är tillämpliga också på ekonomiska utvärderingar av interventioner inom socialtjänsten. Även om hälsa inte alltid ingår som ett viktigt utfall i de utvärderingarna, är metodiken och tankegången likartade.

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

- Sjukdomars och sociala problems påverkan på livskvalitet och kostnader
- Hälsoekonomiska utvärderingar:
  - Systematiska översikter av befintlig litteratur om kostnadseffektivitet (empiriska studier och modeller)
  - Egna kostnadseffektivitetsanalyser
- - Budgetpåverkansanalyser

12.2 SBU:s arbete med hälsoekonomiska utvärderingar

12.2.1 Sjukdomars och sociala problems påverkan på livskvalitet och kostnader

Sjukdom, ohälsa och sociala problem kan beskrivas och mätas utifrån olika perspektiv. Dessa kan vara individens egna (självrapporterad sjuklighet) eller professionens definition baserad på kliniska kriterier (diagnostiserad sjuklighet).

Ett sätt att ge en övergripande beskrivning av sjuklighet och sociala problem är att beräkna de samlade kostnaderna de leder till för samhället. Denna typ av studier brukar kallas för cost-of-illness-studier (COI) [178][179]. Ett annat sätt är att beräkna förlusten i friska år genom att använda mått som kombinerar

livslängd och hälsa, oftast Quality-Adjusted Life Years (QALYs).

Kvalitetsjusterade levnadsår (Quality adjusted life year), QALY

Det rekommenderas ofta att den hälsoekonomiska analysen ska använda kvalitetsjusterade levnadsår (QALY) som effektmått [180][181][182]. QALY mäter både tid (överlevnad) och livskvalitet, det vill säga både livslängd och hälsostatus inklusive effekter av eventuella biverkningar. Livskvalitet mäts på en skala mellan 0 och 1 där 0=död och 1=full hälsa. Exempelvis ger 5 år med en livskvalitet på 0,7 sammanlagt 3,5 QALY (5 x 0,7). Fördelen med QALY 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 livskvalitetsvikter, så kallade QALY-vikter.

QALY-vikter kan skattas med direkta och indirekta metoder, läs mer om dem nedan.

Direkta och indirekta metoder för att skatta QALY-vikter

Direkta metoder för att skatta QALY

De direkta metoderna för att skatta QALYs används för att skatta värdet av olika hälsotillstånd. De vanligaste direkta metoderna är standard gamble (SG) [183], time trade-off (TTO) [184] och visual analogue scale (VAS) [185]. Alla kan användas såväl för att be patienter och brukare 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 markerar hur de värderar ett hälsotillstånd på en linje mellan bästa tänkbara tillstånd och sämsta tänkbara tillstånd.

Indirekta metoder för att skatta QALY

De indirekta metoderna för att skatta QALYs består av ett frågeformulär, ofta kallat livskvalitetsinstrument, som kan kopplas till en värdering, ett värderingssystem (även kallat tariff eller algoritm), som tagits fram med någon av de direkta metoderna. De mest förekommande indirekta instrumenten är EQ-5D [186], SF-6D [187] och HUI-3 [188]. Det finns även andra instrument, till exempel AQoL [189] och särskilda instrument framtagna för barn och ungdomar [190]. Frågeformulären de bygger på ser olika ut och de värderingssystem som används för att omvandla svaren i formulären till QALY-vikter har tagits fram på olika sätt. Även vilken befolkningsgrupp som gjort värderingen skiljer sig åt, där tre grupper förekommer: den allmänna befolkningen, patienter/brukare som värderar sitt eget hälsotillstånd samt experter. Generellt brukar patienter ge högre värden än den allmänna befolkningen [191].

Samhällskostnaden eller påverkan på hälsa och välfärd för olika sjukdomar och sociala problem ger viss information om problemets storlek, men dessa ger inte besked om olika metoders kostnadseffektivitet, och utgör därmed inget beslutsunderlag för fördelning av resurser i samhället [192][193].

12.2.2 Systematisk översikt av hälsoekonomiska studier

Det första steget i SBU:s arbete med att beskriva kostnadseffektivitet är ofta att göra en systematisk översikt över publicerad hälsoekonomisk litteratur. En litteratursökning görs utifrån de söktermer som använts för projektets sökning, och kompletteras med ekonomiska sökord.

Kvaliteten på hälsoekonomiska utvärderingar är beroende av kvaliteten på data och de principer som använts för att beräkna kostnader och effekter. Den ekonomiska utvärderingen kan därför inte bli bättre än vad ingående data möjliggör [193][194][195]. SBU har därför, baserat på tidigare checklistor [193][194][195][196] och erfarenhet utvecklat två egna mallar för kvalitetsgranskning; en för empiriska studier och en för modellstudier. De har

gemensam grund men har anpassats för att bättre fånga de specifika frågor som gäller de olika typerna av studiedesign. Mallarna inkluderar också frågor om överförbarheten till svenska förhållanden och risk för jäv. För att beskriva resultatet av kvalitetsgranskningen anges om studierna, efter en samlad bedömning, är av hög, medel eller låg kvalitet. Läs mer om överförbarhet nedan.

Överförbarhet till svenska förhållanden

Överförbarheten till svenska förhållanden bedöms utifrån hur väl de olika delarna i den hälsoekonomiska analysen stämmer överens med svenska data. Skillnader i organisation, kostnader, dödlighet och livskvalitet samt skillnader i epidemiologiska data påverkar alla resultatet av den hälsoekonomiska analysen [197][198]. Generellt skulle det vara bäst om alla data kunde hämtas från svenska datakällor av god kvalitet [199]. Majoriteten av de hälsoekonomiska analyser som har publicerats är dock genomförda i andra länder, så en viktig del i SBU:s granskning är att bedöma i vilken mån en analys med svenska data skulle ge ett likartat resultat.

12.2.3 Egna analyser

Ofta kan den publicerade litteraturen inte besvara projektets hälsoekonomiska frågeställning. Antalet studier kan vara för få, eller så är resultaten från studier i andra länder inte relevanta för svenska förhållanden. En möjlighet är att istället göra egna analyser, under förutsättning att det går att få fram trovärdiga data på kostnader och effekter. Analyserna kan bli mer eller mindre omfattande, beroende på hur komplex frågan är och hur tillgången på data ser ut.

Ibland räcker det med enbart ett resonemang kring metodens kostnader för att bedöma kostnadseffektiviteten, medan det i andra fall kan bli aktuellt att göra egna modellanalyser. Dessa görs vanligen med utgångspunkt från tillgängliga kliniska studier och anpassas till svenska förhållanden (t.ex. kostnadsdata). Projektgruppens sakkunniga konsulteras också för att bedöma om de uppgifter som använts i kalkylerna är relevanta och korrekta. För att undersöka resultatets osäkerhet bör modellberäkningarna bli föremål för utförlig känslighetsanalys.

12.3 Hälsoekonomiska utvärderingar och kostnadseffektivitet

I hälsoekonomiska utvärderingar jämförs två eller flera alternativa behandlingsmetoder. Man jämför både kostnader och effekter i syfte att klargöra vilken metod som är kostnadseffektiv i jämförelse med det andra alternativet eller alternativen [200]. Om en ny metod har lägre kostnad och bättre effekt än den jämförda metoden så kallas den nya metoden ”dominant” och valet av metod är enkelt ur en hälsoekonomisk synpunkt; välj den nya metoden. Dock är effektivare metoder oftast mer kostnadskrävande. Det finns nio alternativ som kan uppkomma vid en jämförelse och de kan sammanställas i en beslutsmatris. Läs mer om detta nedan.

Beslutsmatris för kostnadseffektivitet

I en beslutsmatris (se Tabell 12.1) visas de nio alternativ som kan uppkomma vid en jämförelse mellan metoder.

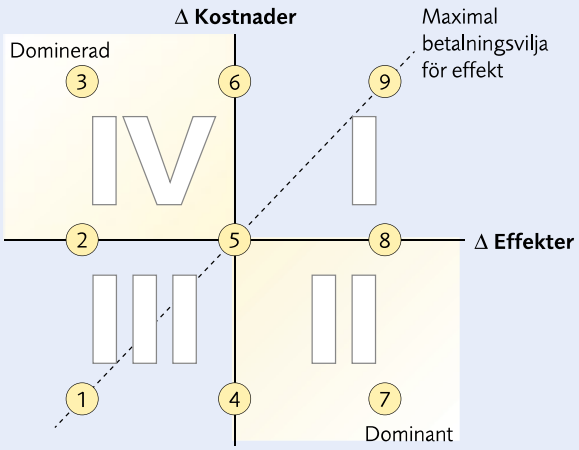
Tabell 12.1 Beslutsmatris för kostnadseffektivitet.

Ny metod jämförs med gammal	Lägre effekt	Lika effekt	Högre 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, ev inkrementell analys

Vid alternativen 2, 3 och 6 är den gamla metoden kostnadseffektiv och behålls. Vid alternativen 4, 7 och 8 gäller motsatsen, det vill säga 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 det ytterligare analyser *eventuellt* för alternativ 1, men *definitivt* för alternativ 9 (se också Figur 12.1).

Resultatet kan även beskrivas i ett så kallat kostnadseffektivitetsplan, där värdena för kvoten placeras i en figur med fyra kvadranter (se Figur 12.1). Då kvadrant IV respektive II ger självskrivna svar (enligt IV dominerar den gamla metoden, enligt II dominerar den nya metoden) fokuseras intresset i allmänhet främst på kvadrant I och III. I dessa återfinns resultatet då den nya metoden medför högre effekt, men också högre kostnad – eller lägre kostnad, men lägre effekt, jämfört med alternativet. Om man vet samhällets maximala betalningsvilja för en effektenhet kan man rita in en gräns för vad som är kostnadseffektivt. Denna gräns går då igenom kvadranterna I och III och alla insatser som har en kostnadseffektivitetskvot till höger om denna linje uppfattas som kostnadseffektiva.

Figur 12.1 Kostnadseffektivitetsplan.



I = Högre kostnad/högre effekt; II = Lägre kostnad/högre effekt
III = Lägre kostnad/lägre effekt; IV = Högre kostnad/lägre effekt jämfört med alternativet

Kostnadseffektivitet är alltså ett relativt begrepp. Ibland kan det mest relevanta alternativet emellertid vara ”ingen behandling”. Det är vanligt att skilja mellan fem olika typer av hälsoekonomiska utvärderingar. Samtliga utvärderingar inkluderar kostnader men skiljer sig åt när det gäller beskrivning och värdering av effekter, se Tabell 12.2.

Tabell 12.2 Olika typer av hälsoekonomiska analysmetoder.

Typ av utvärdering	Effektmått	Hur analysens resultat presenteras
Kostnads-minimeringsanalys (Cost Minimisation Analysis, CMA)	Inget effektmått då effekterna förutsätts vara helt lika	Endast kostnader
Kostnads-konsekvensanalys (Cost Consequences Analysis, CCA)	Flera olika mått på effekter, till exempel antal hemtjänstbesök, förmåga att promenera och anhörigas livskvalitet	Kostnader och effekter, men utan att räkna samman dem i ett mått
Kostnadseffekts-analys (Cost Effectiveness Analysis, CEA)	Fysiska enheter, till exempel levnadsår, antal personer med lyckat resultat, genomsnittlig minskning i riskmarkör	Kostnad per effekt, till exempel per vunnet levnadsår (LYS), per enhets förbättring i depressionsskala
Kostnadsnyttoanalys (Cost Utility Analysis, CUA)	Mått som kombinerar överlevnad och livskvalitet, till exempel QALY	Kostnad per vunnet till exempel QALY
Kostnadsintäkts-analys (Cost Benefit Analysis, CBA)	Olika mått på effekter, till exempel minskad smärta, värderade som intäkter i monetära termer	Nettokostnad

Valet av metod bestäms av frågan, men även av tillgången på relevanta data. Om utvärderingen ska ligga till grund för val mellan två behandlingsmetoder, med samma effektivitet och inga skillnader vad gäller negativa konsekvenser (t.ex. biverkningar), så är det naturligt att nöja sig med en kostnadsminimeringsanalys (CMA). Handlar det om alternativa metoder som främst påverkar dödligheten kan det räcka att göra en kostnadseffektanalys (CEA) med levnadsår som effektmått. Om det däremot rör sig om behandling av kroniska tillstånd som inte är direkt livshotande, är det nödvändigt att även beakta effekterna på livskvalitet, vilket gör kostnadsnyttoanalysen (CUA) till en lämplig metod.

Resultatet från en hälsoekonomisk analys presenteras ofta som en inkrementell kostnadseffektivitetskvot (ICER), vilken är kvoten mellan kostnadsskillnad och effektskillnad.

$$\text{ICER} = \frac{\text{Kostnad A} - \text{Kostnad B}}{\text{Effekt A} - \text{Effekt B}}$$

Kvoten (ICER) anger alltså kostnaden för att uppnå ytterligare en effektenhet (till exempel ett vunnet levnadsår) när man väljer den ena metoden framför den andra.

En metod bedöms som kostnadseffektiv i förhållande till en annan om dess ICER är lägre än samhällets betalningsvilja för en enhet av utfallsmåttet, till exempel en QALY. Gränsen för samhällets betalningsvilja brukar ofta kallas tröskelvärdet. Idealt skulle betalningsviljan för en QALY representera dess alternativkostnad. För att det ska vara motiverat att införa en ny metod måste

den nya metoden producera samma effekt men till en lägre kostnad. Annars innebär det att vi får ut mindre hälsa av ett införande än vad vi redan hade med dagens fördelning av resurser.

För att ovanstående resonemang ska fungera krävs det att vi vet kostnaden per QALY för allt som betalas av samhällets skattemedel, vilket tyvärr är praktiskt omöjligt. Av dessa anledningar antas ofta samhällets betalningsvilja för en QALY uppgå till ett visst tröskelvärde [201][202][203]. I Sverige har det inte satts en exakt gräns för hur mycket en QALY får kosta för att en metod ska anses vara kostnadseffektiv jämfört med alternativbehandlingen. Det finns olika sätt att definiera och ta fram tröskelvärden, vilket leder till en stor variation i publicerade värden [204][205]. En studie från England har skattat kostnaden för att få ytterligare en QALY på marginalen i hälso- och sjukvårdssektorn, vilket i Sverige skulle motsvara ungefär mellan 170 000 och 210 000 kronor [205][206]. För Sverige har en studie skattat ett intervall på 150 000 till 350 000 kronor för vad individer på marginalen är villiga att ge upp i konsumtion för att få ytterligare en QALY [207], medan en annan studie har kommit fram till en siffra på 2,4 miljoner kronor [208].

I Sverige ska prioriteringar inom offentligt finansierad hälso- och sjukvård göras utifrån den etiska plattformen, som omfattar människovärdesprincipen, behovs- och solidaritetsprincipen samt kostnadseffektivitetsprincipen. Detta innebär i praktiken att olika aspekter påverkar betalningsviljan för en QALY, till exempel sjukdomens svårighetsgrad.

12.3.1 Val av perspektiv på analysen

Hälsoekonomiska analyser kan ha ett hälso-och sjukvårdsperspektiv eller ett samhällsperspektiv. Oftast eftersträvas att analysen ska ha 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. Dock kan det ändå vara av intresse att beskriva hur kostnader och effekter fördelar sig på olika intressenter, såsom patient/brukare, landsting, kommun, staten med flera.

12.3.2 Kostnader och besparingar

I en hälsoekonomisk analys ingår både kostnader och kostnadsbesparingar uttryckta i monetära termer. Kostnader uppstår när resurser förbrukas för att ge en viss behandling. Om en åtgärd har en positiv effekt i form av minskad sjuklighet eller sociala problem kan detta också innebära framtida besparingar.

Utifrån ett samhällsperspektiv bör samtliga relevanta kostnader förknippade med de metoder som utvärderas identifieras, kvantifieras och värderas. Ett relevant kostnadsbegrepp inom hälsoekonomin är alternativkostnaden, vilket är värdet av det som kan uppnås av resurserna i bästa alternativa användning. I praktiken används dock marknadspriser eller kostnader härledda ur den offentliga sektorns

kostnadsredovisningar.

Kostnader relaterade till sjukdom, vård och omsorg kan delas in i direkta och indirekta kostnader [209]. Direkta kostnader är den resursförbrukning som uppstår som en direkt följd av vård och behandling såsom personal, facilitet eller kostnader för patienten. Indirekta kostnader beskrivs ofta som de resurser som förloras indirekt på grund av sjukdom eller behandling, till exempel nedsatt arbetsförmåga eller produktionsbortfall. Vilka kostnader som inkluderas beror på vilken typ av metod som utvärderas. Underlag för att beräkna kostnader kan hämtas från svenska register eller statistikkällor.

Svenska register eller statistikkällor för att beräkna kostnader

Socialstyrelsen har 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). Socialstyrelsen har också öppna jämförelser för socialtjänsten i Sverige där uppgifter om bland annat kostnad per brukare och antal hemtjänsttimmar återfinns.

Sveriges Kommuner och Landsting (SKL) har två kostnadsdatabaser: KPP, som innehåller uppgifter om kostnad per patient vid vissa sjukhus, och KPB, kostnad per brukare för vissa kommuners omsorg om äldre och personer med funktionshinder. Regionala priser och ersättningar publiceras också av regionvårdsnämnder. Ytterligare en källa är de nationella kvalitetsregistren som ofta innehåller specifika data om behandlingsinsatser och patientens status.

12.3.3 Att beräkna värdet av produktion

Kostnader för produktionsbortfall uppstår när en individ inte kan arbeta på grund av sjukdom eller för att den får en viss behandling. Även sjuknärvaro, det vill säga att individen arbetar, men som till följd av sin sjukdom eller skada har lägre produktivitet än tidigare, räknas som produktionsbortfall. Individer som inte är i arbetsför ålder inkluderas vanligen inte produktionspåverkan i analysen. Detta har dock kritiserats då ålderspensionärer ofta bidrar med informell produktion, vilket också borde värderas och inkluderas i den hälsoekonomiska analysen [210]. Att inkludera produktionspåverkan i analysen endast för arbetsföra individer kan dessutom anses stå i konflikt med människovärdesprincipen [210][211]. Det har därför rekommenderats att resultatet från hälsoekonomiska analyser presenteras både med och utan produktionspåverkan [200][210], ett förhållningssätt som även SBU rekommenderar.

Det finns två metoder för att skatta värdet av produktion: humankapitalmetoden och friktionskostnadsmetoden [200]. Läs mer om dessa metoder nedan.

Två metoder för att skatta värdet av produktion

Humankapitalmetoden
Med humankapitalmetoden görs värderingen av produktion vanligtvis under antagande att produktionen kan värderas till marknadspris, det vill säga lön plus arbetsgivar- och sociala avgifter.

Friktionskostnadsmetoden
Med friktionskostnadsmetoden görs en värdering av den tid (med tillhörande kostnad) som går innan en tidigare arbetslös individ fullt ut kan ersätta en person [212][213].

12.3.4 Modellanalyser

En modell syftar till att belysa ett beslutsproblem utifrån bästa tillgängliga information, inte att ersätta empiriska studier. I modellanalyser används en mängd olika uppgifter som har samlats in tidigare, ofta kallad sekundärdata, tillsammans med primärdata från prövningar. Det är främst i situationer när kostnader och effekter till följd av åtgärder påverkas över en längre tid än vad som har kunnat studeras i prövningen som modeller tillämpas vid hälsoekonomisk utvärdering. Dessutom är det ofta aktuellt vid följande situationer [214]:

- Då relevanta kliniska utvärderingar saknas eller inte inkluderar data på kostnader och QALYs.
- För att extrapolera från intermediära utfallsmått, exempelvis från blodtryck till hjärtinfarkt.
- Då det av etiska skäl är omöjligt att genomföra kontrollerade kliniska prövningar.
- Då kostnaderna för att genomföra tillräckligt stora empiriska studier är orimligt höga i förhållande till det potentiella värdet av ytterligare information som kan vinnas.
- Att kostnader som beräknats inom ramen för kliniska prövningar inte är realistiska eller att de inte är relevanta för svenska förhållanden.

De vanligaste teknikerna vid modellanalyser inom hälsoekonomin är så kallade beslutsträd och Markov-modeller [214]. Principerna för dessa två metoder är i stort lika, men ett beslutsträd visar en sekvens av händelser under en bestämd tidsperiod. Det har på senare tid även blivit vanligare att använda sig av händelsestyrda modeller (eng. discrete event simulation, DES) [215].

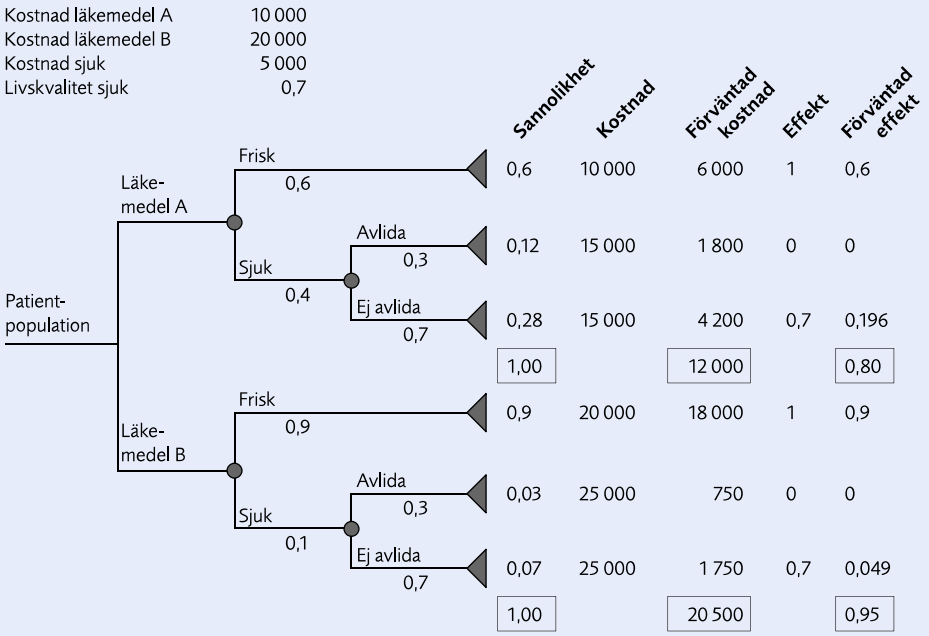
Modellanalyser: beslutsträd, Markov-modeller och Discrete event stimulation (DES)

Denna teknik är lämplig vid utvärdering rörande sjukdomar eller problem av mer akut karaktär med ett händelseförlopp som är begränsat till en relativt kort tidsperiod.

I Figur 12.2 jämförs två alternativa läkemedelsbehandlingar (A och B) med hjälp av ett beslutsträd. Modellen består 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 kostnadseffektivitetskvoten (ICER), det vill säga merkostnaden per effektenhet om man väljer läkemedel B istället för A, blir (20 500–12 000)/(0,95–0,80) = 56 667 kronor.

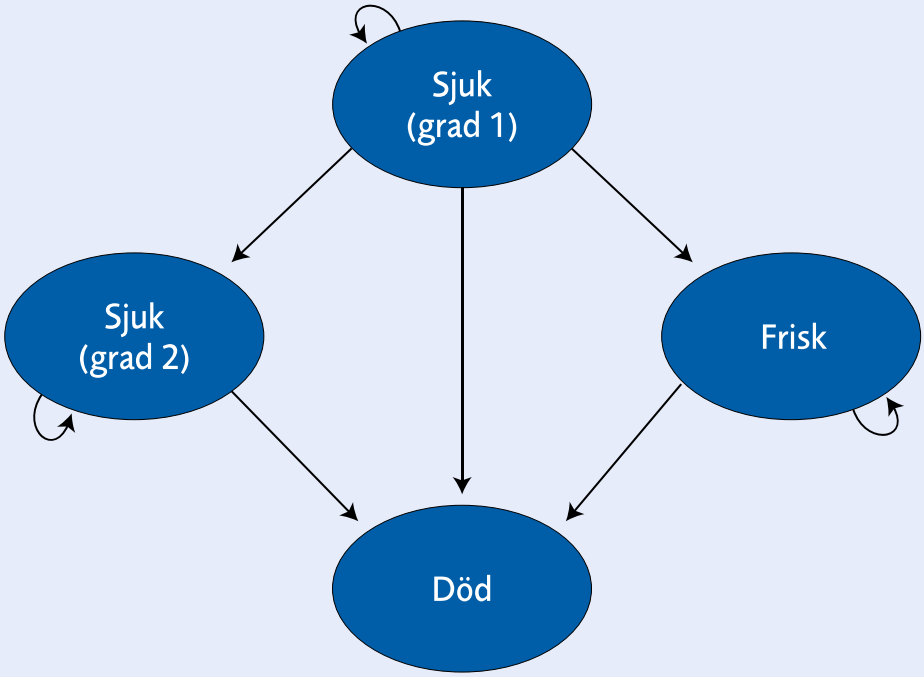
Figur 12.2 Beslutsträd. Exempel på jämförelse av två alternativa läkemedel (A och B)



Markov-modeller

Dessa modeller är uppbyggda kring ömsesidigt uteslutande tillstånd, se Figur 12.3. Varje tillstånd är förknippat med en viss kostnad och en viss QALY-vikt. Modellerna innehåller alltid ett initialt tillstånd, exempelvis sjuk i en viss sjukdom eller tonåring med ökad risk för kriminalitet, och ett slutligt tillstånd, vanligtvis död. Pilarna i figuren representerar övergångsannolikheter, det vill säga risker, för förflyttningar mellan de olika hälsotillstånden. Dessa risker kan i moderna Markov-modeller (via så kallad mikrosimulering) tillåtas variera över tid, exempelvis öka med åldern på patienterna. Markov-modellen är mer användbar för analys av beslutsproblem som avser lång tid, till exempel behandling av kroniska problem, och är därför den vanligaste skattningsmodellen.

Figur 12.3 Exempel på Markov-modell



Discrete event simulation, DES
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 såsom att en patient insjuknar, ett läkarbesök eller att en viss behandling påbörjas. Flera olika händelser kan ske samtidigt och var och en av dessa händelser kan i sin tur få konsekvenser i form av till exempel 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 [216], men den uppfattningen delas inte av alla [215].

12.3.5 Känslighetsanalyser

Vid hälsoekonomiska utvärderingar är det viktigt att göra känslighetsanalyser [209] för att beskriva osäkerheten i resultatet. Att göra en känslighetsanalys innebär att man varierar en eller flera variabler i analysen för att undersöka vad som händer med analysens resultat. Om resultatet förändras mycket, till exempel så att kostnadseffektivitetskvoten förändras till att bli högre än tröskelvärdet för samhällets betalningsvilja, säger man att resultatet är känsligt för variabeln.

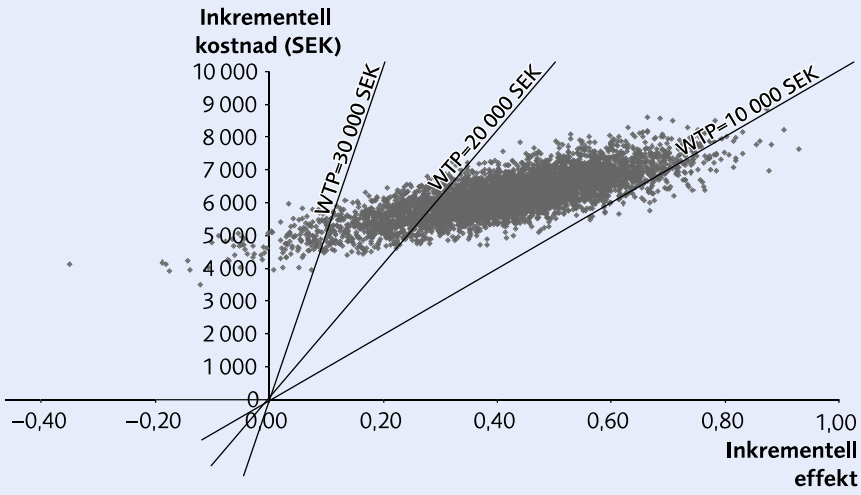
Det finns olika typer av osäkerhetsanalyser, varav ”bootstrapping ” är ett exempel. Denna metod appliceras på data från empiriska studier som kostnader och effekter vars resultat sedan presenteras i ett kostnadseffektivitetsplan [192]. I modeller brukar probabilistisk känslighetsanalys (eng. probabilistic sensitivity analysis, PSA) tillämpas [214], vilket innebär att osäkerheten kring modellens variabler analyseras.

Känslighetsanalyser: bootstrapping och probabilistisk känslighetsanalys (PSA)

Bootstrapping
Innebär att man skattar datauppgifters statistiska fördelning genom att använda de datauppgifter man har. Vid empiriska studier kan man använda de två gruppernas data på kostnader och effekter för att beräkna ett stort antal kostnadseffektivitetskvoter. Detta görs genom att slumpmässigt dra individdata ur de två grupperna och beräkna skillnaden. Detta gör man ett stort antal gånger, ofta minst 1 000 gånger, och varje gång läggs individdatan tillbaka så att den kan dras igen. Bootstrapping ger alltså ett stort antal olika kostnadseffektivitetskvoter, och ett slags konfidensintervall kan anges genom att beräkna mellan vilka värden 95 procent av dragningarna återfinns. Motsvarande metod kan användas för att beskriva den empiriska, icke-parametriska, fördelningen av kostnader och effekter separat, både från behandlingsstudier och modellskattningar.

Probabilistisk känslighetsanalys (PSA)
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 att beräkna en förväntad kostnad per effekt. I varje körning dras ett värde från varje variabelfördelning och ett resultat beräknas. I Figur 12.4 illustreras resultatet av en modell som körts 5 000 gånger. Linjerna i figuren anger olika nivåer för betalningsviljan 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.

Figur 12.4 Kostnadseffektivitetsplan med probabilistisk känslighetsanalys.



SEK = *Svenska kronor;* **WTP** = *Willingness to pay (maximal betalningsvilja för effekt)*

12.4 Analys av budgetpåverkan

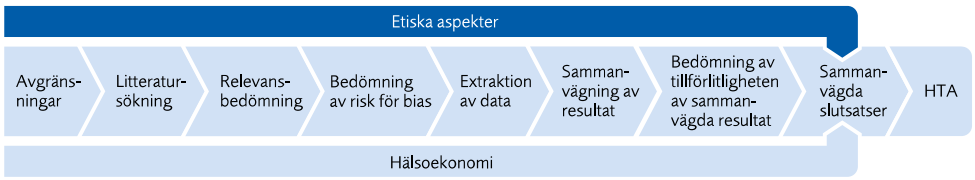
För att underlätta för dem som ska finansiera införandet av en viss metod kan kostnadseffektivitetsanalyserna kompletteras med en budgetpåverkananalys (eng. budget impact analysis). Med hjälp av analysen beskriver man hur en viss eller flera budgetar påverkas vid införandet av en metod och vilka konsekvenser man kan förvänta. Den utvärderar alltså inte 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 publicerat riktlinjer för budgetpåverkananalyser [217], vilka du hittar här.

12.5 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. Tanken med hälsoekonomiska utvärderingar är att de ska användas som stöd vid beslutsfattande. Av den anledningen görs ofta andra analyser och statistiska test än till exempel de som genomförs för att fastställa en medicinsk åtgärds kliniska effekt.

Resultat från modeller som bygger på ett flertal olika källor och antaganden ska inte tolkas som evidens utan som en skattning av en metods påverkan på kostnader och hälsoeffekter. Däremot är det, utifrån SBU:s synpunkt, viktigt att de effektmått som modellen bygger på är statistiskt säkerställda. Hälsoekonomiska utfallsmått i randomiserade kontrollerade studier (t.ex. antal vårddagar) kan evidensgraderas precis som de medicinska utfallsmåtten men kostnadseffektivitetskvoten låter sig inte evidensgraderas på vanligt sätt eftersom den består av en sammanslagning av ett flertal olika utfallsmått [194].

13. Etiska aspekter



13.1 En del av beslutsunderlaget

Etiska aspekter på metoder (terapeutiska, stödjande eller diagnostiska) kan, tillsammans med både medicinska, hälsoekonomiska och sociala aspekter, stå för en viktig del i beslutfattares underlag vid beslut om införande, fortsatt användning eller utmönstring av metoder i hälso- och sjukvården, eller inom socialtjänsten. Till viss del är betydelsen av etiska bedömningar begränsade av tvingande lagar kring olika verksamheter, främst inom socialtjänsten. Samtidigt kräver andra lagar inom hälso- och sjukvård att man gör etiska avvägningar vid införandet av vissa nya metoder, som ”kan ha betydelse för människovärde och integritet” (HSL 5 kap 3§), eller så utgör lagrummet stöd och ramverk för etiska värderingar (prop 1996/97:60, HSL 3 kap 1§ och 4 kap 1§).

13.1.1 Arbetet med etiska aspekter

Projektgruppen bör diskutera arbetet med etiska aspekter tidigt under projektprocessen. I projektplanen bör man beskriva både omfattning och inriktning av detta arbete, och specificera om det finns behov av litteratursökning efter studier kring etiska aspekter. Beroende på vilken metod som utvärderas kan det ibland räcka med en kortare diskussion av etiska aspekter medan det i andra fall behövs en mer omfattande etisk analys. För att identifiera viktiga etiska frågeställningar, intressentkonflikter och olika problemområden kan det vara värdefullt att tidigt i projektarbetet involvera företrädare för patient- och brukarorganisationer, anhörigorganisationer och berörda professioner. Projektgruppen bör också tidigt i processen överväga om det behövs en etikexpert, en mer omfattande etisk diskussion eller analys. I vissa fall kan ett samarbete med Statens medicinsk-etiska råd (Smer) bli aktuellt. Formerna för detta samarbete bör då tydliggöras innan samarbetet startar (Vem äger slutliga utformningen av analysen? Hur ska analysen presenteras i relation till huvudrapporten? etc.).

Processen för arbetet med etiska aspekter beskrivs närmare i de etiska vägledningar, för att identifiera och beskriva etiska aspekter, som SBU tagit fram. Läs mer om dem nedan.

13.2 Identifiering av etiska aspekter

För att underlätta arbetet med att identifiera och beskriva etiska aspekter på utvärderade metoder har SBU utarbetat vägledningar för att identifiera etiska aspekter vid utvärdering av metoder i hälso- och sjukvården respektive

socialtjänsten. Dessa vägledningar är tänkta att användas som stöd för att undvika att viktiga etiska aspekter glöms bort. Dock bör endast de aspekter som är aktuella för metoden tas upp och beskrivas i rapportens etikkapitel. I vägledningarna understryker man vikten av att en initial diskussion hålls inom projektgruppen för att identifiera relevanta etiska aspekter *innan* man går igenom frågelistorna, för att säkerställa att varken vägledningens eller projektgruppens intuitivt identifierade aspekter missas. Exempel på aspekter som lyfts i vägledningarna är åtgärdens påverkan på jämlikhet, rättvisa, autonomi, integritet och strukturella faktorer med etiska implikationer. Det är av största betydelse att också lyfta de etiska problem som kan uppstå på grund av utvärderingens resultat.

13.2.1 Speciella förutsättningar för det sociala området

Socialtjänsten i Sverige arbetar under särskilda förutsättningar som kan ha etisk betydelse. Viktiga förutsättningar är den tydliga lokala politiska styrningen och kopplingen till den lagtolkning som finns inom det sociala området. Lagstödet för prioriteringar mellan olika gruppers behov är mindre tydligt än i hälso- och sjukvården. Detta kan påverka möjligheten att väga in sådant som storleken på individens behov eller hänsyn till kostnadseffektivitet, vilka är centrala begrepp inom hälso- och sjukvårdens prioriteringar.

13.2.2 Identifiering av mål- och intressekonflikter

En viktig del i arbetet med att beskriva etiska aspekter på utvärderade metoder är att identifiera de olika grupper som berörs, och de eventuella etiska målkonflikter eller intressemotsättningar som finns inom respektive grupp, eller mellan olika grupper. Det kan exempelvis vara patienter/brukare, olika professioner, anhöriga, andra patientgrupper som drabbas av alternativkostnaden samt medborgarna. SBU:s roll är i allmänhet inte att klargöra vilka intressen som har företräde, utan snarare att beskriva hur själva konflikten ser ut och vilka intressen som behöver balanseras. Den så kallade aktörsmodellen är ett sätt att strukturera arbetet och den beskrivs närmare i Smer:s handbok. I denna beskrivs också andra etiska principer och begrepp som kan användas som stöd i arbetet.

13.3 Prioriteringsetik

I rapportens etikkapitel bör projektgruppen också diskutera resultaten från den hälsoekonomiska utvärderingen, som är en ytterligare del inom SBU:s utvärderingsarbete. Resultaten bör diskuteras i relation till den etiska värdegrund (plattform) som gäller för prioriteringar av metoder i hälso- och sjukvård (se fråga 8 i hälso- och sjukvårdsvägledningen).

SBU ska ta fram ett allsidigt beslutsunderlag men det ligger *inte* inom SBU:s uppdrag att prioritera metoder utifrån dessa principer.

13.4 PRISMA -E för utvärdering av jämlikhet och rättvisa

I de fall där arbetet med en SBU-rapport fokuserar på grupper som riskerar att behandlas orättvist, på interventioner som syftar till att påverka ojämlik tillgång till hälso- och sjukvård, eller när de interventioner som rapporten behandlar har betydande implikationer på rättvisa, kan PRISMA-E (Equity) statement, med sitt tillhörande extensionsdokument, vara ett användbart stöd för att ta med dessa aspekter i arbetet. Stödet kan också användas för att bedöma hur andra översikter har hanterat dessa aspekter.

13.5 Forskningsetiska frågor

Att etiskt kontroversiell forskning har använts då man tagit fram kunskap om den aktuella metoden som SBU utvärderar, utgör som regel inte ett etiskt dilemma för beslutet om metoden sedan kan användas eller inte ute i verksamheten. Men, om det saknas kunskap om metodens effekt och/eller säkerhet är det viktigt att identifiera om det skulle innebära etiska dilemman eller forskningsetiska problem att ta fram sådan kunskap, som medför att forskningen kan vara svår att genomföra. Det är i sådana fall viktigt att tydliggöra etiska konsekvenser av olika alternativa sätt att hantera denna kunskapsbrist (se fråga 2 i etikvägledningen för hälso- och sjukvård). När man beskriver forskningsetiska frågeställningar bör man också grunda diskussionen på etikprövningslagen och Helsingforsdeklarationen.

SBU har också utarbetat en etisk vägledning för forskningsfinansiärer vid prioritering av forskningsprojekt i relation till identifierade kunskapsluckor. Etiska problem kan spela stor roll för vilka vetenskapliga kunskapsluckor i hälso- och sjukvården som är särskilt viktiga att forska på. Frågorna i vägledningen är tänkta som ett stöd för att identifiera och reflektera över etiska aspekter vid prioritering av forskningsfrågor.

13.6 Professionsetiska riktlinjer

Det finns en rad andra etikpolicys hos olika professionsföreningar som också kan användas som stöd i arbetet, bland annat hos Läkarförbundet, Sjuksköterskeföreningen och World Medical Association.

14. Referenser

1. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 2009;339:b2535.
2. McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, the P-DTAG, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. JAMA 2018;319:388-96.
3. Tong A, Flemming K, McInnes E, Oliver S, Craig J. Enhancing transparency in reporting the synthesis of qualitative research: ENTREQ. BMC Med Res Methodol 2012;12:181.
4. Flay BR, Biglan A, Boruch RF, Castro FG, Gottfredson D, Kellam S, et al. Standards of evidence: criteria for efficacy, effectiveness and dissemination. Prev Sci 2005;6:151-75.
5. Ferreira-Gonzalez 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;60:651-7; discussion 658-62.
6. Williamson P, Clarke M. The COMET (Core Outcome Measures in Effectiveness Trials) Initiative: Its Role in Improving Cochrane Reviews. Cochrane Database Syst Rev 2012:ED000041.
7. Bernal JL, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. Int J Epidemiol 2017;46:348-355.
8. Penfold RB, Zhang F. Use of interrupted time series analysis in evaluating health care quality improvements. Acad Pediatr 2013;13:S38-44.
9. Ramsay CR, Matowe L, Grilli R, Grimshaw JM, Thomas RE. Interrupted time series designs in health technology assessment: lessons from two systematic reviews of behavior change strategies. Int J Technol Assess Health Care 2003;19:613-23.
10. SBU. Program för att förebygga psykisk ohälsa hos barn. En systematisk litteraturöversikt. Stockholm: Statens beredning för medicinsk utvärdering (SBU); 2010. SBU-rapport nr 202. ISBN 978-91-85413-38-6.
11. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making 1991;11:88-94.
12. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006;332:1089-92.
13. Moriarty AS, Gilbody S, McMillan D, Manea L. Screening and case finding for major depressive disorder using the Patient Health Questionnaire (PHQ-9): a meta-analysis. Gen Hosp Psychiatry 2015;37:567-76.
14. 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.

15. 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.
16. Leeflang MM, Rutjes AW, Reitsma JB, Hooft L, Bossuyt PM. Variation of a test's sensitivity and specificity with disease prevalence. CMAJ 2013;185:E537-44.
17. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29-36.
18. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol 2010;5:1315-6.
19. SBU. Risk- och behovsbedömning av ungdomar avseende återfall i våld och annan kriminalitet. (SBU 2019).
20. Halligan S, Altman DG, Mallett S. Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. Eur Radiol 2015;25:932-9.
21. SBU. Riskbedömningar inom psykiatrin - kan våld i samhället förutsägas?. En systematisk litteraturöversikt. Stockholm: Statens beredning för medicinsk utvärdering (SBU); 2005. SBU-rapport nr 175. ISBN 91-85413-03-8.
22. SBU. Diagnostik och uppföljning av förstämningssyndrom. En systematisk litteraturöversikt. Stockholm: Statens beredning för medicinsk utvärdering (SBU); 2012. SBU-rapport nr 212. ISBN 978-91-85413-52-2.
23. Fischer JE, Bachmann LM, Jaeschke R. A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. Intensive Care Med 2003;29:1043-51.
24. Booth A, Noyes J, Flemming K, Gerhardus A, Wahlster P, van der Wilt GJ, et al. Structured methodology review identified seven (RETREAT) criteria for selecting qualitative evidence synthesis approaches. J Clin Epidemiol 2018;99:41-52.
25. Howell Major C, Savin-Baden M. An Introduction to Qualitative Research Synthesis. Managing the Information Explosion in Social Science Research. Routledge 2010.
26. Patton MQ. Qualitative Research & Evaluation methods. 3 edition. Sage Publications. Thousand Oaks. California; 2002.
27. Trulsson U, Engstrand P, Berggren U, Nannmark U, Brånemark P-I. Edentulousness and oral rehabilitation: experiences from the patient´s perspective. Eur J Oral Sci 2002;110:417-424.
28. Ferguson KM. Exploring family environment characteristics and multiple abuse experiences among homeless youth. J Interpers Violence 2009;24:1875-91.
29. Granskär M, Höglund-Nielsen B. (red). Tillämpad kvalitativ forskning inom hälso- och sjukvård. Lund, Studentlitteratur; 2008.
30. Gannon M, Dowling M. Nurses' experience of loss on the death of older persons in long-term residential care: findings from an interpretative phenomenological study. Int J Older People Nurs 2012;7:243–52.

31. Avby G, Nilsen P, Abrandt Dahlgren M. Ways of understanding evidence-based practice in social work: a qualitative study. Br J Soc Work 2014;44:1366-83.
32. Rapport F. Exploring the beliefs and experiences of potential egg share donors. J Adv Nurs 2003;43:28-42.
33. Lauver LS. The Lived Experience of Foster Parents of Children With Special Needs Living in Rural Areas. J Pediatr Nurs 2010;25:289-98.
34. Holloway I. (ed). Qualitative research in health care. Maidenhead, Open University Press; 2005.
35. Gustafsson M, Kristensson J, Holst G, Willman A, Bohman D. Case managers for older persons with multi-morbidity and their everyday work – a focused ethnography. BMC Health Serv Res 2013;13:496.
36. Lislerud Smebye K, Kirkevold M. The influence of relationships on personhood in dementia care: a qualitative, hermeneutic study. BMC Nurs 2013;12:29.
37. Ellermann CR. Influences on the mental health of Children Placed in Foster Care. Fam Community Health 2007;30(2 Suppl):S23-32.
38. Fern E. Developing social work practice through engaging practitioners in action research. Qual Soc Work 2010;11:156-173.
39. Twigg RC. The unknown soldiers of foster care: foster care as loss for the foster parents' own children. Smith Coll Stud Soc Work 1994:297-312.
40. SBU. Tandförluster. En systematisk litteraturöversikt. Stockholm: Statens beredning för medicinsk utvärdering (SBU); 2010. SBU-rapport nr 204. ISBN 978-91-85413-40-9.
41. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M, et al. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ 2008;337:a1655.
42. Noyes J, Booth A, Cargo M, Fleming K, Harden A, Harris J, et al. Chapter 21: Qualitative evidence. https://training.cochrane.org/handbook/current/chapter-21#section-21-13. In: Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, et al., editors. Cochrane Handbook for Systematic Reviews of Interventions. Cochrane Training; 2019.
43. Noyes J, Hendry M, Booth A, Chandler J, Lewin S, Glenton C, et al. Current use was established and Cochrane guidance on selection of social theories for systematic reviews of complex interventions was developed. J Clin Epidemiol 2016;75:78-92.
44. Witting M, Boere-Boonekamp MM, Fleuren MAH, Sakkers RJB, Ijzerman MJ. Determinants of parental satisfaction with ultrasound hip screening in child health care. Journal of child health care : for professionals working with children in the hospital and community 2012;16:178-189.
45. SCIE. Systematic research reviews: Guidelines. Social Care Institute for Excellence (SCIE); 2013. Avaliable from: https://www.scie.org.uk/publications/researchresources/rr01.pdf.
46. EUnetHTA. Process of information retrieval for systematic reviews and

health technology assessments on clinical effectiveness: EUnetHTA (European Network for Health Technology Assessment); 2017 [cited 2019 October].
47. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane Handbook for Systematic Reviews of Interventions version 6.0 (updated July 2019). Cochrane, 2019. Available from https://www.training.cochrane.org/handbook.
48. Kugley S, Wade A, Thomas J, Mahood Q, Jørgensen AMK, Hammerstrøm K, et al. Searching for studies: a guide to information retrieval for Campbell systematic reviews. In. The Campbell Collaboration, Oslo, Norway; 2017.
49. Lefebvre C, Glanville J, Briscoe S, Littlewood A, Marshall C, Metzendorf MI, et al. Technical Supplement to Chapter 4: Searching for and selecting studies. In: Higgins JPT, Thomas J, Chandler J, Cumpston MS, Li T, Page MJ, et al., editors. Cochrane Handbook for Systematic Reviews of Interventions Version 6. Cochrane; 2019.
50. Atkinson KM, Koenka AC, Sanchez CE, Moshontz H, Cooper H. Reporting standards for literature searches and report inclusion criteria: making research syntheses more transparent and easy to replicate. Res Synth Methods 2015;6:87-95.
51. Cooper C, Booth A, Varley-Campbell J, Britten N, Garside R. Defining the process to literature searching in systematic reviews: a literature review of guidance and supporting studies. BMC Med Res Methodol 2018;18:85.
52. Metzendorf MI, Featherstone RM. Ensuring quality as the basis of evidence synthesis: leveraging information specialists' knowledge, skills, and expertise. Cochrane Database Syst Rev 2018;4:Ed000125.
53. Rethlefsen ML, Farrell AM, Osterhaus Trzasko LC, Brigham TJ. Librarian co-autors correlated with higher quality reported search strategies in general internal medicine systematic reviews. Journal of Clinical Epidemilogy 2015:999-1000.
54. Kelly MP, Noyes J, Kane RL, Chang C, Uhl S, Robinson KA, et al. AHRQ series on complex intervention systematic reviews-paper 2: defining complexity, formulating scope, and questions. Journal of clinical epidemiology 2017;90:11-8.
55. Egger M, Zellweger-Zahner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomised controlled trials published in English and German. London, England, Lancet 1997.
56. Morrison A, Polisena J, Husereau D, Moulton K, Clark M, Fiander M, et al. The effect of English-language restriction on systematic review-based meta-analyses: a systematic review of empirical studies. International journal of technology assessment in health care 2012;28:138-44.
57. Hartling L, Featherstone RM, Nuspl M, Shave K, Dryden DM, Vandermeer B. Grey literature in systematic reviews: a cross-sectional study of the contribution of non-English reports, unpublished studies and dissertations to the results of meta-analyses in child-relevant reviews. BMC medical research methodology 2017;17:64.

58. Nussbaumer-Streit B, Klerings I, Dobrescu AI, Persad E, Stevens A, Garritty C. Excluding non-English publications from evidence-syntheses did not change conclusions: a meta-epidemiological study. Journal of clinical epidemiology 2020;118:42-54.
59. Harbour J, Fraser C, Lefebvre C, Glanville J, Beale S, Boachie C, et al. Reporting methodological search filter performance comparisons: a literature review. Health information and libraries journal 2014;31:176-94.
60. Sampson M, Tetzlaff J, Urquhart C. Precision of healthcare systematic review searches in a cross-sectional sample. Research synthesis methods 2011;2:119-2.
61. Booth A, Papaioannou D, Sutton A. Systematic approaches to a successful literature review. London, UK, Sage; 2012.
62. Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 [uppdaterad 2011].Tillgänglig från: www.handbook.cochrane.org [Hämtad 2019-05-11]: The Cochrane Collaboration; 2011.
63. Arber M, Glanville J, Isojarvi J, Baragula E, Edwards M, Shaw A, et al. Which databases should be used to identify studies for systematic reviews of economic evaluations? International journal of technology assessment in health care 2018;34:547-54.
64. Pitt C, Goodman C, Hanson K. Economic Evaluation in Global Perspective: A Bibliometric Analysis of the Recent Literature. Health economics 2016;25 Suppl 1:9-28.
65. Glanville J, Paisley S. Searching for evidence for cost-effectiveness decisions. In: Shemilt I, Mugford M, Vale L, Marsh K, Donaldson C, editors. Evidence-based decisions and economics: health care, social welfare, education and criminal justice. 2 ed. Chichester: Wiley; 2010.
66. Glanville J, Fleetwood K, Yellowlees A, Kaunelis D, Mensinkai S. Development and Testing of Search Filters to Identify Economic Evaluations in MEDLINE and EMBASE. Ottawa: Canadian Agency for Drugs and Technologies in Health (CADTH); 2009 [cited 2019-11-20]. Available from: https://www.cadth.ca/media/pdf/H0490_Search_Filters_for_Economic_Evaluations_mg_e.pdf.
67. Glanville J, Kaunelis D, Mensinkai S. How well do search filters perform in identifying economic evaluations in MEDLINE and EMBASE. International journal of technology assessment in health care 2009;25:522-9.
68. Hemminki E. Study of information submitted by drug companies to licensing authorities. BMJ 1980;280:833-6.
69. Eyding D, Lelgemann M, Grouven U, Harter M, Kromp M, Kaiser T, et al. Reboxetine for acute treatment of major depression: systematic review and meta-analysis of published and unpublished placebo and selective serotonin reuptake inhibitor controlled trials. BMJ (Clinical research ed) 2010;341:c4737.
70. Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective

publication of antidepressant trials and its influence on apparent efficacy. New England Journal of Medicine 2008;358:252-60.
71. Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, et al. Dissemination and publication of research findings: an updated review of related biases. Health technology assessment 2010;14:iii, ix-xi, 1-193.
72. Jefferson T, Doshi P, Boutron I, Golder S, Heneghan C, Hodkinson A, et al. When to include clinical study reports and regulatory documents in systematic reviews. BMJ evidence-based medicine 2018;23:210-7.
73. Jefferson T, Jones MA, Doshi P, Del Mar CB, Hama R, Thompson MJ, et al. Neuraminidase inhibitors for preventing and treating influenza in healthy adults and children. 2014;4:Cd008965.
74. Agency for Healthcare Research and Quality (AHRQ). Methods guide for effectiveness and comparative effectiveness reviews: AHRQ publication no. 10(14)-EHC063-EF 2014 [cited 2019-11-13]. Available from: https://effectivehealthcare.ahrq.gov/sites/default/files/pdf/cer-methods-guide_overview.pdf.
75. The Campbell Collaboration. Campbell systematic reviews: policies and guidelines. Version 1.4. 2019 [cited 2019 September 22]. Available from: https://wol-prod-cdn.literatumonline.com/pb-assets/assets/18911803/Campbell%20Policies%20and%20Guidelines%20v4.pdf.
76. IQWIG. General methods. Version-5-0: IQWIG Institute for Quality and Efficiency in Health Care (IQWiG); 2017.
77. Isojarvi J, Wood H, Lefebvre C, Glanville J. Challenges of identifying unpublished data from clinical trials: Getting the best out of clinical trials registers and other novel sources. Research synthesis methods. February 2018.
78. Scherer RW, Meerpohl JJ, Pfeifer N, Schmucker C, Schwarzer G, von Elm E. Full publication of results initially presented in abstracts. The Cochrane database of systematic reviews; 2018;11:Mr000005.
79. Li G, Abbade LPF, Nwosu I, Jin Y, Leenus A, Maaz M, et al. A scoping review of comparisons between abstracts and full reports in primary biomedical research. BMC Medical Research Methodology 2017;17:181.
80. Scherer RW, Saldanha IJ. How should systematic reviewers handle conference abstracts? A view from the trenches. Systematic reviews 2019;8:264.
81. Baudard M, Yavchitz A, Ravaud P, Perrodeau E, Boutron I. Impact of searching clinical trial registries in systematic reviews of pharmaceutical treatments: methodological systematic review and reanalysis of meta-analyses. BMJ (Clinical research ed) 2017;356:j448.
82. Knelangen M, Hausner E, Metzendorf MI, Sturtz S, Waffenschmidt S. Trial registry searches for randomized controlled trials of new drugs required registry-specific adaptation to achieve adequate sensitivity. Journal of clinical epidemiology 2018;94:69-75.
83. Glanville JM, Duffy S, McCool R, Varley D. Searching ClinicalTrials.gov

and the International Clinical Trials Registry Platform to inform systematic reviews: what are the optimal search approaches? Journal of the Medical Library Association : JMLA 2014;102:177-83.
84. Schmucker CM, Blumle A, Schell LK, Schwarzer G, Oeller P, Cabrera L, et al. Systematic review finds that study data not published in full text articles have unclear impact on meta-analyses results in medical research. PloS one 2017;12:e0176210.
85. Halfpenny NJ, Quigley JM, Thompson JC, Scott DA. Value and usability of unpublished data sources for systematic reviews and network meta-analyses. Evidence-based medicine 2016;21:208-13.
86. Brolund A. Söka grå litteratur till systematiska översikter: Vad säger ett urval metodböcker och nyare studier? SBU praxis? Stockholm: SBU, Statens beredning för medicinsk och social utvärdering; 2018.
87. Adams J, Hillier-Brown FC, Moore HJ, Lake AA, Araujo-Soares V, White M, et al. Searching and synthesising 'grey literature' and 'grey information' in public health: critical reflections on three case studies. Systematic reviews 2016;5:164.
88. Mahood Q, Van Eerd D, Irvin E. Searching for grey literature for systematic reviews: challenges and benefits. Research synthesis methods 2014;5:221-34.
89. Booth A, Noyes J, Flemming K, Gerhardus A, Wahlster P, van der Wilt GJ, et al. Guidance on choosing qualitative evidence synthesis methods for use in health technology assessments of complex interventions. https://www.integrate-hta.eu/wp-content/uploads/2016/02/Guidance-on-choosing-qualitative-evidence-synthesis-methods-for-use-in-HTA-of-complex-interventions.pdf. 2016.
90. Booth A. Searching for qualitative research for inclusion in systematic reviews: a structured methodological review. Systematic review 2016;5:74.
91. Frandsen TF, Gildberg FA, Tingleff EB. Searching for qualitative health research required several databases and alternative search strategies: a study of coverage in bibliographic databases. Journal of clinical epidemiology 2019;114:118-24.
92. Papaioannou D, Sutton A, Carroll C, Booth A, Wong R. Literature searching for social science systematic reviews: consideration of a range of search techniques. Health information and libraries journal 2010;27:114-22.
93. Harris JL, Booth A, Cargo M, Hannes K, Harden A, Flemming K, et al. Cochrane Qualitative and Implementation Methods Group guidance series-paper 2: methods for question formulation, searching, and protocol development for qualitative evidence synthesis. Journal of clinical epidemilogy 2018;97:39-48.
94. NICE. Developing NICE guidelines: the manual: process and methods. In. NICE (National Institute for Health and Care Excellence); 2018.
95. Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. Syst Rev 2017;6:245.
96. Mayo-Wilson E, Li T, Fusco N, Dickersin K, investigators M. Practical

guidance for using multiple data sources in systematic reviews and meta-analyses (with examples from the MUDS study). Res Synth Methods 2018;9:2-12.
97. Stevinson C, Lawlor DA. Searching multiple databases for systematic reviews: added value or diminishing returns? Complement Ther Med 2004;12:228-32.
98. Cooper C, Booth A, Britten N, Garside R. A comparison of results of empirical studies of supplementary search techniques and recommendations in review methodology handbooks: a methodological review. Syst Rev 2017;6:234.
99. Horsley T, Dingwall O, Sampson M. Checking reference lists to find additional studies for systematic reviews. Cochrane Database Syst Rev 2011:MR000026.
100. Cooper C, Lovell R, Husk K, Booth A, Garside R. Supplementary search methods were more effective and offered better value than bibliographic database searching: A case study from public health and environmental enhancement. Res Synth Methods 2018;9:195-223.
101. Biocic M, Fidahic M, Puljak L. Reproducibility of search strategies of non-Cochrane systematic reviews published in anaesthesiology journals is suboptimal: primary methodological study. Br J Anaesth 2019;122:e79-e81.
102. Koffel JB, Rethlefsen ML. Reproducibility of Search Strategies Is Poor in Systematic Reviews Published in High-Impact Pediatrics, Cardiology and Surgery Journals: A Cross-Sectional Study. PLoS One 2016;11:e0163309.
103. Sampson M, McGowan J, Tetzlaff J, Cogo E, Moher D. No consensus exists on search reporting methods for systematic reviews. J Clin Epidemiol 2008;61:748-54.
104. Schulz KF, Altman DG, Moher D, Group C. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. Ann Intern Med 2010;152:726-32.
105. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015;351:h5527.
106. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007;19:349-57.
107. Munthe-Kaas HM, Glenton C, Booth A, Noyes J, Lewin S. Systematic mapping of existing tools to appraise methodological strengths and limitations of qualitative research: first stage in the development of the CAMELOT tool. BMC Med Res Methodol 2019;19:113.
108. Higgins JPT, Sterne JAC, Page MJ, Hróbjartsson A, Boutron I, Reeves B, et al. A revised tool for assessing risk of bias in randomized trials In: Chandler J, McKenzie J, Boutron I, Welch V (editors). Cochrane Methods. Cochrane Database of Systematic Reviews 2016, Issue 10 (Suppl 1). dx.doi.org/10.1002/14651858.CD201601.
109. Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan

M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016;355:i4919.
110. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Group Q-S. A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. J Clin Epidemiol 2013;66:1093-104.
111. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529-36.
112. Lincoln YS, Guba EG. But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation. no. 30. In: Williams DD, editor. Naturalistic Evaluation; 1986. p 73-84.
113. Graneheim UH, Lundman B. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. Nurse Educ Today 2004;24:105-12.
114. Munthe-Kaas H, Bohren MA, Glenton C, Lewin S, Noyes J, Tuncalp O, et al. Applying GRADE-CERQual to qualitative evidence synthesis findings-paper 3: how to assess methodological limitations. Implement Sci 2018;13:9.
115. Noyes J, Booth A, Flemming K, Garside R, Harden A, Lewin S, et al. Cochrane Qualitative and Implementation Methods Group guidance series-paper 3: methods for assessing methodological limitations, data extraction and synthesis, and confidence in synthesized qualitative findings. J Clin Epidemiol 2018;97:49-58.
116. Noyes J, Booth A, Lewin S, Carlsen B, Glenton C, Colvin CJ, et al. Applying GRADE-CERQual to qualitative evidence synthesis findings-paper 6: how to assess relevance of the data. Implement Sci 2018;13:4.
117. Malterud K, Siersma VD, Guassora AD. Sample Size in Qualitative Interview Studies: Guided by Information Power. Qual Health Res 2016;26:1753-1760.
118. 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.
119. Whiting P, Savovic J, Higgins JP, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol 2016;69:225-34.
120. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996;17:1-12.
121. Scottish Intercollegiate Guidelines Network (SIGN). Healthcare Improvement Scotland. Critical appraisal notes and checklists (2014).
122. Dwan K, Gamble C, Williamson PR, Kirkham JJ, Reporting Bias G. Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. PLoS One 2013;8:e66844.
123. 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.
124. Sterne JA, Egger M, Smith GD. Systematic reviews in health care: Investigating and dealing with publication and other biases in meta-analysis. BMJ 2001;323:101-5.
125. van Aert RCM, Wicherts JM, van Assen M. Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis. PLoS One 2019;14:e0215052.
126. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. Introduction to Meta-Analysis. John Wiley & Sons, Ltd. 2009.
127. Guyatt GH, Thorlund K, Oxman AD, Walter SD, Patrick D, Furukawa TA, et al. GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles-continuous outcomes. J Clin Epidemiol 2013;66:173-83.
128. Thorlund K, Walter SD, Johnston BC, Furukawa TA, Guyatt GH. Pooling health-related quality of life outcomes in meta-analysis-a tutorial and review of methods for enhancing interpretability. Res Synth Methods 2011;2:188-203.
129. SBU. Endometrios – Diagnostik, behandling och bemötande: en systematisk översikt och utvärdering av medicinska, hälsoekonomiska, sociala och etiska aspekter. Stockholm: Statens beredning för medicinsk och social utvärdering (SBU); 2018. SBU-rapport nr 277. ISBN 978-91-88437-19-8.
130. 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.
131. Bond GR, Witheridge TF, Dincin J, Wasmer D, Webb J, De Graaf-Kaser R. Assertive community treatment for frequent users of psychiatric hospitals in a large city: a controlled study. Am J Community Psychol 1990;18:865-91.
132. 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.
133. 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.
134. 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.
135. 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.
136. 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.

137. 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.
138. 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. Med Care 1998;36:40-53.
139. Clarke GN, Herinckx HA, Kinney RF, Paulson RI, Cutler DL, Lewis K, et al. 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.
140. 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 1.0. The Cochrane Collaboration, 2010. Available from: http://srdta.cochrane.org/.
141. 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.
142. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med 2001;20:2865-84.
143. Takwoingi Y, Guo B, Riley RD, Deeks JJ. Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. Stat Methods Med Res 2017;26:1896-1911.
144. Arends LR, Hamza TH, van Houwelingen JC, Heijenbrok-Kal MH, Hunink MG, Stijnen T. Bivariate random effects meta-analysis of ROC curves. Med Decis Making 2008;28:621-38.
145. Kester AD, Buntinx F. Meta-analysis of ROC curves. Med Decis Making 2000;20:430-9.
146. Noblit GW, Hare RD. Meta-Ethnography: Synthesizing Qualitative studies. Sage Publications, Newbury Park. 1988.
147. Thorne S, Jensen L, Kearney MH, Noblit G, Sandelowski M. Qualitative metasynthesis: reflections on methodological orientation and ideological agenda. Qual Health Res 2004;14:1342-65.
148. Finfgeld-Connett D. Use of content analysis to conduct knowledge-building and theory-generating qualitative systematic reviews. Qualitative Research 2014;14:341-352.
149. SBU. Rehabilitering för vuxna med traumatisk hjärnskada. En systematisk översikt och utvärdering av medicinska, ekonomiska, sociala och etiska aspekter. Stockholm: Statens beredning för medicinsk och social utvärdering (SBU); 2019. SBU-rapport nr 304. ISBN 978-91-88437-46-4.
150. Hannes K, Lockwood C. Pragmatism as the philosophical foundation for the Joanna Briggs meta-aggregative approach to qualitative evidence synthesis. J Adv Nurs 2011;67:1632-42.
151. Lockwood C, Munn Z, Porritt K. Qualitative research synthesis:

methodological guidance for systematic reviewers utilizing meta-aggregation. Int J Evid Based Healthc 2015;13:179-87.
152. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol 2008;8:45.
153. SBU. Myalgisk encefalomyelit och kroniskt trötthetssyndrom (ME/CFS). En systematisk översikt. Stockholm: Statens beredning för medicinsk och social utvärdering (SBU); 2018. SBU-rapport nr 295. ISBN 978-91-88437-37-2.
154. SBU. Läkemedelsbehandling av vanliga smärttillstånd hos äldre personer. Effekter, biverkningar samt upplevelser av vård. Stockholm: Statens beredning för medicinsk och social utvärdering (SBU); 2020. SBU-rapport nr 315. ISBN 978-91-88437-57-0.
155. 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.
156. Schünemann H, Brożek J, Gordon G, Oxman A. GRADE Handbook. Introduction to GRADE Handbook.Handbook for grading the qualityof evidence and the strength of recommendations using the GRADE approach. Updated October 2013. Available from: https://gdt.gradepro.org/app/handbook/handbook.html.
157. Schunemann HJ, Cuello C, Akl EA, Mustafa RA, Meerpohl JJ, Thayer K, et al. GRADE guidelines: 18. How ROBINS-I and other tools to assess risk of bias in nonrandomized studies should be used to rate the certainty of a body of evidence. J Clin Epidemiol 2019;111:105-114.
158. Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). J Clin Epidemiol 2011;64:407-15.
159. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. J Clin Epidemiol 2011;64:1294-302.
160. Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence--imprecision. J Clin Epidemiol 2011;64:1283-93.
161. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence--indirectness. J Clin Epidemiol 2011;64:1303-10.
162. Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. J Clin Epidemiol 2011;64:1277-82.
163. Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol 2011;64:1311-6.
164. Guyatt GH, Oxman AD, Santesso N, Helfand M, Vist G, Kunz R, et al. GRADE guidelines: 12. Preparing summary of findings tables-binary outcomes. J Clin Epidemiol 2013;66:158-72.

165. Lewin S, Glenton C, Munthe-Kaas H, Carlsen B, Colvin CJ, Gulmezoglu M, et al. Using qualitative evidence in decision making for health and social interventions: an approach to assess confidence in findings from qualitative evidence syntheses (GRADE-CERQual). PLoS Med 2015;12:e1001895.
166. Colvin CJ, Garside R, Wainwright M, Munthe-Kaas H, Glenton C, Bohren MA, et al. Applying GRADE-CERQual to qualitative evidence synthesis findings-paper 4: how to assess coherence. Implement Sci 2018;13:13.
167. Glenton C, Carlsen B, Lewin S, Munthe-Kaas H, Colvin CJ, Tuncalp O, et al. Applying GRADE-CERQual to qualitative evidence synthesis findings-paper 5: how to assess adequacy of data. Implement Sci 2018;13:14.
168. 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.
169. 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). PLoS One 2007;2:e1350.
170. SBU. Tidig koordinerad utskrivning och fortsatt rehabilitering i hemmiljö för äldre efter stroke. En systematisk litteraturöversikt. Stockholm: Statens beredning för medicinsk utvärdering (SBU); 2015. SBU-rapport nr 234. ISBN 978-91-85413-77-5.
171. SBU. Rehabilitering av äldre personer med höftfrakturer – interdisciplinära team. En systematisk litteraturöversikt. Stockholm: Statens beredning för medicinsk utvärdering (SBU); 2015. SBU-rapport nr 235. ISBN 978-91-85413-79-9.
172. SBU. Ljusbehandling och systemisk behandling av psoriasis. En systematisk översikt och utvärdering av medicinska, hälsoekonomiska och etiska aspekter. Stockholm: Statens beredning för medicinsk och social utvärdering (SBU); 2018. SBU-rapport nr 278. ISBN 978-91-88437-20-4.
173. SBU. Att förebygga missbruk av alkohol, droger och spel hos barn och unga. Stockholm: Statens beredning för medicinsk och social utvärdering (SBU); 2015. SBU-rapport nr 243. ISBN 978-91-85413-87-4.
174. White CM, Ip S, McPheeters M, Carey TS, Chou R, Lohr KN, et al. Using Existing Systematic Reviews To Replace De Novo Processes in Conducting Comparative Effectiveness Reviews. In: Methods Guide for Effectiveness and Comparative Effectiveness Reviews. Rockville (MD); 2008.
175. Whitlock EP, Lin JS, Chou R, Shekelle P, Robinson KA. Using existing systematic reviews in complex systematic reviews. Ann Intern Med 2008;148:776-82.
176. Newdick C. Who should we treat? Rights, Rationing, and Resources in the NHS, Oxford university press; 2005.
177. Stevens A, Milne R, Burls A. Health technology assessment: history and demand. J Public Health Med 2003;25:98-101.
178. 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.
179. Rice DP. Estimating the cost of illness. Am J Public Health Nations Health 1967;57:424-40.
180. National Institute for Health and Clinical excellence (NICE). Guide to the methods of technology appraisal. 2008.
181. International society for pharmacoeconomics and outcomes research. Pharmacoeconomic guidelines around the world. ISPOR.; 2008.
182. Tandvårds- och läkemedelsförmånsverket (TLV). Tandvårds- och läkemedelsförmånsverkets allmänna råd om ekonomiska utvärderingar. TLVAR 2017:1. Tandvårds- och läkemedelsförmånsverket (TLV); 2017.
183. von Neumann J, Morgenstern O. Theory of Games and Economic Behavior., Princeton University Press; 1944.
184. Torrance GW, Thomas WH, Sackett DL. A utility maximization model for evaluation of health care programs. Health Serv Res 1972;7:118-33.
185. 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.
186. The EuroQol Group. EuroQol--a new facility for the measurement of health-related quality of life. Health Policy 1990;16:199-208.
187. 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.
188. 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.
189. Mihalopoulos C, Chen G, Iezzi A, Khan MA, Richardson J. Assessing outcomes for cost-utility analysis in depression: comparison of five multi-attribute utility instruments with two depression-specific outcome measures. Br J Psychiatry 2014;205:390-7.
190. Chen G, Ratcliffe J. A Review of the Development and Application of Generic Multi-Attribute Utility Instruments for Paediatric Populations. Pharmacoeconomics 2015;33:1013-28.
191. Bernfort L. QALY som effektmått inom vården. Möjligheter och begränsningar. CMT Rapport 2012:2. Linköpings Universitet. 2012.
192. Byford S, Torgerson DJ, Raftery J. Economic note: cost of illness studies. BMJ 2000;320:1335.
193. Drummond M. Cost-of-illness studies: a major headache? Pharmacoeconomics 1992;2:1-4.
194. Brunetti M. Chapter 10: Grading economic evidence. In: Schemilt I, Mugford M, Vale L, Marsch K, Donaldson C, editors. Evidence-based decisions and economics: Health care, social welfare, education and criminal justice. Oxford: Wiley-Blackwell; 2010.
195. 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.
196. 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.
197. Drummond M, Barbieri M, Cook J, Glick HA, Lis J, Malik F, et al. Transferability of economic evaluations across jurisdictions: ISPOR Good Research Practices Task Force report. Value Health 2009;12:409-18.
198. Mulligan J-A, Fox-Rushby J. Transferring cost-effectiveness data across space and time. In: Fox-Rushby J, Cairns J, editors. Economic evaluation.: Open University Press; 2005.
199. 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.
200. Drummond M, Sculpher M, Torrance GW, O´Brien B, Stoddart G. Methods for the economic evaluation of health care programmes Oxford University Press; 2005.
201. 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.
202. McCabe C, Claxton K, Culyer AJ. The NICE cost-effectiveness threshold: what it is and what that means. Pharmacoeconomics 2008;26:733-44.
203. Rawlins MD, Culyer AJ. National Institute for Clinical Excellence and its value judgments. BMJ 2004;329:224-7.
204. Claxton K, Martin S, Soares M, Rice N, Spackman E, Hinde S, et al. Methods for the estimation of the National Institute for Health and Care Excellence cost-effectiveness threshold. Health Technol Assess 2015;19:1-503, v-vi.
205. Henriksson M, Siverskog J, Johannesen K, Eriksson T. Tröskelvärden och kostnadseffektivitet - innebörde och implikationer för ekonomiska utvärderingar och beslutsfattande i hälso- och sjukvården. CMT Rapport 2018:3. Linköpings Universitet. 2018.
206. Woods B, Revill P, Sculpher M, Claxton K. Country-level Cost-Effectiveness Thresholds: Initial Estimates and the Need for Further Research. CHE Research Paper 109. Centre for Health Economics, University of York, UK. 2015.
207. Olofsson S, Persson U, Hultkrantz L, Gerdtham UG. Betalningsviljan för att minska risken för icke-dödliga och dödliga skador i samband med vägtrafikolyckor – en pilotstudie med jämförelse av CV och kedje-ansats. IHE Rapport 2016:8. Institutet för hälso- och sjukvårdsekonomi. 2016.
208. Robinson A, Gyrd-Hansen D, Bacon P, Baker R, Pennington M, Donaldson C, et al. Estimating a WTP-based value of a QALY: the 'chained' approach. Soc Sci Med 2013;92:92-104.
209. Gold M, Siegel J, Russell L, Weinstein MC. Cost-Effectiveness in Health and Medicine., Oxford University Press; 1996.
210. Sahlén K-G, Löfgren C, Lindholm L. Är det lönsamt med prevention efter 65? : ålderns betydelse i hälsoekonomiska utvärderingsmetoder :

förebyggande hembesök i Nordmaling. Stockholm, Statens folkhälsoinstitut; 2006.
211. Sculpher M. The role and estimation of productivity costs in economic evaluation. In: Drummond M, McGuire A, editors. Economic evaluation in health care: merging theory with practice.: Oxford University Press; 2001.
212. Johannesson M, Karlsson G. The friction cost method: a comment. J Health Econ 1997;16:249-55; discussion 257-9.
213. 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.
214. Briggs A, Claxton K, Sculpher M. Decision modelling for health economic evaluation Oxford University Press; 2006.
215. Caro JJ, Moller J, Getsios D. Discrete event simulation: the preferred technique for health economic evaluations? Value Health 2010;13:1056-60.
216. 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.
217. 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.

Vanligt förekommande termer på SBU, med definitioner

Term:	**systematisk översikt**
Synonym:	(rekommenderas inte: systematisk litteraturöversikt)
Beskrivning:	sammanställning av resultat från sådana studier som med systematiska och explicita metoder har identifierats, valts ut och bedömts kritiskt och som avser en specifikt formulerad fråga
Engelska termer:	*systematic review, systematic overview*
Kommentar:	Innehåller ofta men inte alltid en metaanalys. Ett viktigt syfte med systematiken är att minimera snedvridning.
Term:	**metaanalys**
Synonym:	–
Beskrivning:	
Engelska termer:	meta-analysis
Kommentar:	–
Term:	**nätverks-metaanalys**
Synonym:	–
Beskrivning:	statistisk analysmetod (typ av metaanalys) som jämför två eller flera interventioner genom att kombinera resultat från jämförande primärstudier och indirekta jämförelser
Engelska termer:	*network meta-analysis, NMA*
Kommentar:	–
Term:	**evidens**
Synonym:	(rekommenderas inte: samlat vetenskapligt bevisläge, samlat vetenskapligt underlag)
Beskrivning:	forskningsresultat som är systematiskt sökta, relevans- och kvalitetsgranskade och sammanvägda
Engelska termer:	*evidence*
Kommentar:	I ett medicinskt och socialt vetenskapligt sammanhang.Vid avsaknad av evidens föreligger en vetenskaplig kunskapslucka. Evidensen avser alltid en tydligt formulerad fråga, till exempel vilken effekt en åtgärd har för en viss population eller hur ett samband ser ut i ett specifikt sammanhang.
Term:	**tillförlitlighet**
Synonym:	evidensstyrka
Beskrivning:	giltigheten hos ett systematiskt sökt, relevans- och kvalitetsgranskat och sammanvägt forskningsresultat (evidens) bedömt utifrån: • hur stor risken är för systematiska fel i studierna (engelska: bias, snedvridning), • hur mycket studierna motsäger varandra (engelska: inconsistency, bristande samstämmighet), • i vilken grad som de studerade förhållandena skiljer sig från den aktuella frågan (engelska: indirectness, bristande överförbarhet), • hur stor den statistiska osäkerheten är (engelska: imprecision, bristande precision) samt

- hur stor risken är för snedvriden publicering av studier och resultat (engelska: publication bias).

Engelska termer:	*certainty* (enligt GRADE)
Kommentar:	Gäller medicinsk och social vetenskaplig forskning.
Term:	**systematiskt fel**
Synonym:	bias, snedvridning
Beskrivning:	ett resultatfel i forskningsprocessen som uppstått i en studies upplägg, genomförande, effektbedömning, publikation eller annan hantering av resultaten, och som inte beror på slumpen
Engelska termer:	*bias*
Kommentar:	–
Term:	**originalartikel**
Synonym:	–
Beskrivning:	(när det är själva publikationen som åsyftas): artikel där (vetenskapliga) resultat, åsikter eller synpunkter framläggs för första gången
Engelska termer:	*original article*
Kommentar:	ordet originalstudie om det är själva undersökningen som åsyftas
Term:	**evidenskartläggning**
Synonym:	kartläggning (rekommenderas inte: systematisk kartläggning)
Beskrivning:	inventering av befintliga systematiska översikter (samt i vissa fall enskilda studier) inom ett område, i syfte att påvisa dels forskningsresultat som är systematiskt sökta, relevans- och kvalitetsgranskade och sammanvägda, dels vetenskapliga kunskapsluckor
Engelska termer:	*evidence map, evidence-and-gap map, systematic map*
Kommentar:	Resulterar i en evidenskarta. Skilj från scoping review
Term:	**kostnadseffektivitetsanalys**
Synonym:	–
Beskrivning:	hälsoekonomisk analysmetod där kostnader och effekter av två eller fler insatser jämförs och vars resultat presenteras som en inkrementell kostnadseffektkvot (ICER)
Engelska termer:	*cost-effectiveness analysis*
Kommentar:	ICER beräknas enligt formeln: (kostnad A – kostnad B)/(effekt A – effekt B). Det finns olika typer av kostnadseffektivitetsanalys som använder olika utfallsmått
Term:	**confounder**
Synonym:	förväxlingsfaktor, störfaktor
Beskrivning:	faktor som är kopplad till både en intervention (eller exponering eller omständighet) och effekt, och som därför antingen kan dölja specifika samband mellan exponering och effekt eller skapa skenbara samband mellan exponering och effekt
Engelska termer:	*confounder, confounding factor*
Kommentar:	Termen förväxlingsfaktor bedöms som något tydligare än störfaktor. Vanliga exempel

	i studier av hälsa är ålder, kön och olika former av missbruk. Confound-ing är den vilseledning som uppstår på grund av en confound-er när data från en studie tolkas, och är inte synonymt med confounder
Term:	**prioriterade utfall**
Synonym:	(rekommenderas inte: centrala utfall)
Beskrivning:	överenskommet minimiurval av utfall (och utfallsmått) vilka prioriterats som särskilt viktiga att mäta och rapportera i studier av effekter av insatser vid ett specifikt tillstånd
Engelska termer:	*core outcome set, COS*
Kommentar:	Såväl utfall som utfallsmått bör specificeras, det vill säga både vad som ska mätas och hur. Detta tas fram i projekt där en eller flera intressentgrupper deltar i en gemensam prioritering, och där samverkan med patienter och brukare är en särskilt viktig del i processen.
Term:	**HTA**
Synonym:	(rekommenderas inte: metodutvärdering inom hälso- och sjukvården)
Beskrivning:	tvärvetenskaplig process som använder specificerade utvärderingsmetoder för att bedöma värdet av en hälso- och sjukvårdsåtgärd i något stadium av dess livscykel, och som syftar till att ta fram ett beslutsunderlag som främjar likvärdig och effektiv hälso- och sjukvård av hög kvalitet
Engelska termer:	*HTA, health technology assessment*
Kommentar:	• en åtgärd kan syfta till att förebygga, undersöka, behandla, befrämja hälsa, rehabilitera eller till att organisera hälso- och sjukvård; exempel är testprocedurer, medicintekniska produkter, läkemedel, vacciner, ingrepp, terapier, program och system • processen är formaliserad, systematisk, transparent och använder optimal utvärderingsmetodik för att ta fram bästa tillgängliga evidens • olika aspekter av en åtgärds värde utvärderas genom analys av dess avsedda och icke avsedda konsekvenser genom jämförelse med befintliga åtgärdsalternativ. Ofta innefattar detta klinisk effektivitet och säkerhet; kostnader och ekonomiska effekter; etiska, sociala, kulturella och juridiska aspekter; organisatoriska och miljömässiga aspekter; betydelse för patienten, närstående, vårdare och befolkningen. Det samlade värdet kan vara olika beroende på vilket perspektiv som har anlagts, vilka berörda parter som har tagits med och vilket beslutssammanhang som underlaget avser. • HTA kan genomföras i olika skeden av en åtgärds livscykel – såväl tidigt, det vill säga före, under och efter godkännande- och införandefas, som sent, då användning av åtgärden överges