REVIDERAD 2017

9 Sammanvägning av resultat

Vid utvärderingar av interventioner och metoder inom hälso- och sjukvård samt socialtjänst gäller det att bedöma om ett alternativ är effektmässigt överlägset ett annat för ett givet tillstånd. Om det finns flera studier, behöver studieresultaten vägas samman. De sammanvägda resultaten kan ingå i en evidensprofil (se Kapitel 10 om GRADE) och därefter fungera som en del i ett beslutsunderlag inom evidensbaserad medicin [1]. Om sammanvägningen görs med hjälp av statistiska metoder kallas den för metaanalys; om statistiska metoder inte används brukar man tala om narrativa sammanvägningar. Metaanalyser används oftast avseende randomiserade studier (RCT). De förekommer dock även vid sammanvägningar av andra typer av studier, till exempel inom diagnostik. Syftet med detta kapitel är att ge en orientering om följande:

- vad det innebär att göra en metaanalys
- problem som uppkommer vid en metaanalys samt strategier för att hantera problemen
- publikationsbias
- bristande samstämmighet (heterogenitet)
- random effects model
- metaanalys vid observationsstudier
- nätverksmetaanalys

Alla resultat har inte samma tyngd

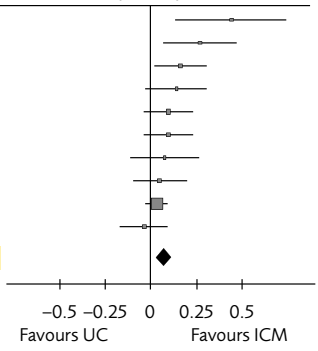
Metaanalys innebär att man räknar fram ensammanvägd effektstorlek från flera enskilda studieresultat. för att skatta en ”sann” effekt. Alla enskilda resultat har dock normalt inte samma tyngd i sammanvägningen. Intuitivt kan man tycka att små studier borde väga mindre än stora studier vid sammanvägningen.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

Ett vanligt sätt att åskådliggöra metaanalysen är en så kallad forest plot (skogsdiagram). Denna innehåller bland annat skattade effektstorlekar för varje studie, en sammanvägd effektstorlek samt konfidensintervall för såväl de enskilda effekterna som för den sammanvägda effekten. I Figur 9.1 visas en forest plot, med resultaten av en intervention för hemlösa personer med psykisk funktionsnedsättning och mer eller mindre grava missbruksproblem [3–12]. Interventionen består av ett program kallat intensive case management (ICM) medan kontrollalternativet består av standardvård (UC för usual care). Effektmåttet är riskskillnad (risk difference)[1]. Riskskillnad 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 9.1
Exempel på metaanalys (forest plot) – intensive case management (ICM) mot standardvård (UC).

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

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



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

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

För Lehman och medarbetares studie är resultatet följande: riskskillnaden är 14 procentenheter, alltså 14 procentenheter fler i interventionsgruppen än kontrollgruppen hade ett eget stabilt boende vid 12-månadersuppföljningen. Konfidensintervallet, från –3 till 32 procentenheter, överlappar emellertid 0-linjen. Detta innebär att skillnaden ligger inom den statistiska felmarginalen. Resultatet är med andra ord inte statistiskt signifikant. Diamanten (romboiden) längst ner visar den sammanvägda effekten samt konfidensintervallet för den sammanvägda effekten: en riskskillnad på 7 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 medan det resultat som väger tyngst har presenterats i en studie av Conrad och medarbetare (41,2 procent). Notera att ett resultat väger tyngre ju kortare konfidensintervallet är. Detta beror på att ju större standardfelet är, desto längre blir konfidensintervallet.

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

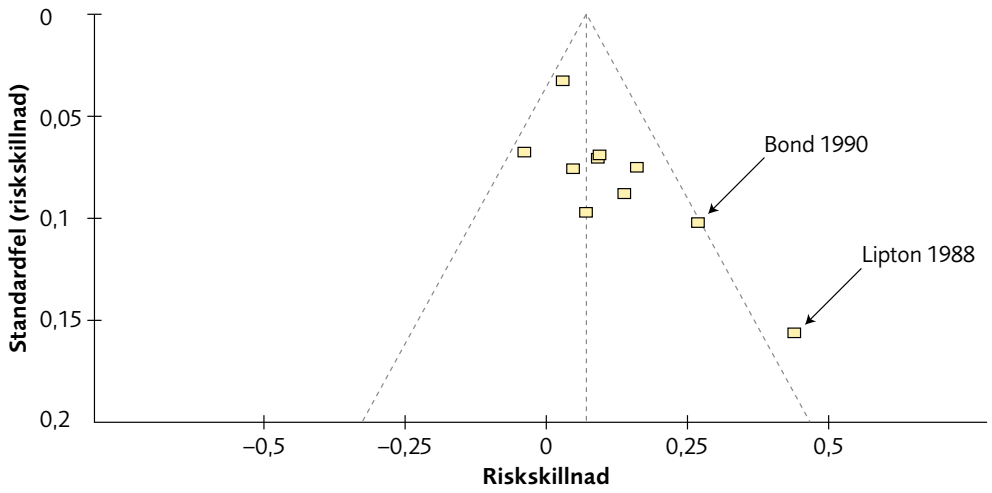
Det finns emellertid några problem som gör att den sammanvägda effekten i Figur 9.1 inte alltid är en tillförlitlig skattning av den ”sanna” effekten. För det första kan det vara så att de resultat som ingår i metaanalysen inte utgör ett representativt urval på grund av ett problem som kallas publikationsbias. 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 [13] och kan ta sig uttryck i såväl en över- som en underskattning av den ”sanna” effekten. I följande avsnitt kommer vi att visa hur metaanalys kan användas för att hantera sådana problem, först publikationsbias och därefter heterogenitet.

Publikationsbias och funnel plots

I Figur 9.2 har resultaten från Figur 9.1 gjorts om till en funnel plot (tratt-diagram). Effektstorleken visas på den horisontella axeln medan spridningen (standardfelet) visas på den vertikala axeln. Observera att den vertikala axelns

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

värden är omvända så att ju högre upp på axeln ett resultat finns, desto mindre är spridningen. Kvadraten längst ner till höger visar resultatet från Lipton och medarbetare med en effekt på 44 procentenheter och med den största spridningen av alla ingående studier. Den streckade triangeln är en hjälp för att visuellt kunna tolka resultatet. Den lodräta mittenlinjen visar var den sammanvägda effekten på 7 procentenheter ligger.



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

Modellen bygger bland annat på två antaganden: (a) att resultat från stora studier (med liten spridning) är lättare att publicera än resultat från små och (b) att resultat med en stor effekt till förmån för den utvärderade interventionen är lättare att publicera än resultat som inte är signifikanta eller som talar emot interventionen [2,14]. Publiceringssvårigheten kan ta sig uttryck i att de små icke-positiva resultaten aldrig blir publicerade, att publiceringen tar längre tid eller att publicering sker i tidskrifter som inte indexeras i referensdatabaser (och kan därmed vara svåra att hitta). Medvetenheten om dessa publiceringsproblem kan även leda till en selektiv rapportering inom varje enskild studie på så sätt att man endast rapporterar de statistiskt signifikanta resultat som talar för interventionen och undviker att rapportera övriga resultat. Ibland talar man om rapporteringsbias, något som inte riktigt är samma sak som publikationsbias. Rapporteringsbias innebär en tendentiös rapportering inom en och samma studie, alltså en benägenhet att endast rapportera de resultat som stödjer interventionen. Om rapporteringsbias är mer vanligt inom småstudier än inom stora, tar sig detta samma uttryck som publiceringsbias. Det bör även nämnas att det kan finnas ekonomiska intressen bakom denna typ av selektiv rapportering om de som utvärderar interventionen kan ha eget intresse av att den framstår som effektiv.

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

betyder att riskskillnaden på 7 procentenheter kan vara en överskattning av den ”sanna” effekten.

För att få en bild av hur mycket effekten överskattas kan man plocka bort de mest extrema resultaten (trim) till förmån för interventionen och därefter räkna fram en ny effektstorlek. För att det sammanvägda konfidensintervallets längd inte ska överskattas kan nya hypotetiska resultat läggas till (fill). Detta sätt att hantera publikationsbias har utvecklats till en statistisk metod kallad trim and fill där man med hjälp av en iterativ process räknar fram ett resultat där publikationbias har hanterats [2]. Om till exempel resultatet från Lipton och medarbetares studie tas bort, förändras inte den skattade effekten på 7 procentenheter, men konfidensintervallets övre gräns minskar från 0,12 till 0,11. Om även Bond och medarbetares studie tas bort, minskar effekten till 6 procentenheter och konfidensintervallet går från 0,02 till 0,10; resultatet är alltjämt statistiskt signifikant. Ovanstående metodologiska övningar kan ge en bild av resultatens konsistens och hur stort ett publikationsbias skulle kunna vara i detta exempel.

Bristande samstämmighet kan tydliggöras och undersökas

I detta avsnitt kommer vi att visa hur problemet med heterogenitet kan hanteras med hjälp av metaanalys [2,13]. Även om alla resultat utom ett i Figur 9.1 uppvisar en positiv effekt är inte resultaten samstämmiga. Exempelvis varierar effektstorleken en hel del, från 44 procentenheter (Lipton och medarbetare) till minus 4 procentenheter (Clarke och medarbetare). 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 χ2-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 I^2, 44 procent i fallet ovan. Annorlunda uttryckt, I^2 utgör andelen av den totala variansen som förklaras av att det finns reella skillnader i effektstorlekar studierna emellan. Enligt en tumregel brukar I^2 benämnas på följande sätt: låg heterogenitet = 0,25, måttlig heterogenitet = 0,50 och hög heterogenitet = 0,75 [2].

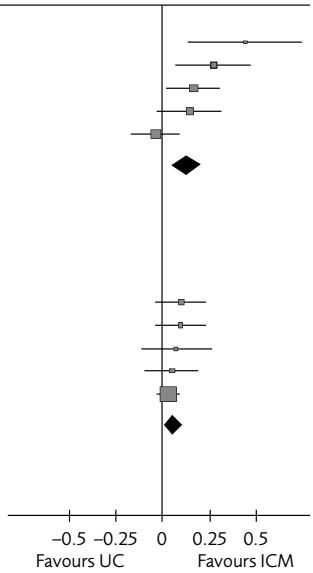
Anta att de olika resultaten bygger på studier som är mycket lika varandra avseende interventioner, kontrollvillkor, utvärderingsdesign och effektmått. Anta vidare att populationerna varierar från ett resultat till ett annat, men att positiva effekter trots detta uppvisar stor samstämmighet. Under sådana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

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

I Figur 9.3 har resultaten delats upp i två subgrupper men någon total sammanvägning har inte gjorts. Med denna gruppindelning framgår att det inte finns någon heterogenitet inom SA/MISA-gruppen medan den 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 9.3
Subgrupper – psykisk funktionsnedsättning och drogmissbruk, intensive case management (ICM) mot standardvård (UC).

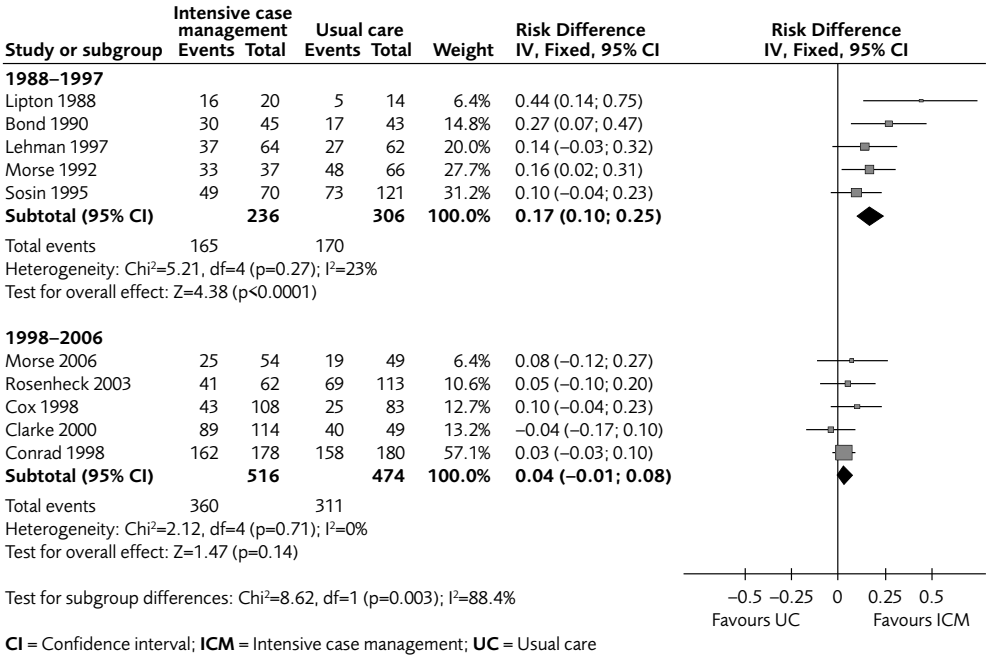
Study or subgroup	Intensive case management Events	Total	Usual care Events	Total	Weight	Risk Difference IV, Fixed, 95% CI
MI						
Lipton 1988	16	20	5	14	6.2%	0.44 (0.14; 0.75)
Bond 1990	30	45	17	43	14.4%	0.27 (0.07; 0.47)
Morse 1992	33	37	48	66	26.9%	0.16 (0.02; 0.31)
Lehman 1997	37	64	27	62	19.4%	0.14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



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

Bakom heterogeniteten kan det finnas ett metodologiskt problem. Detta problem kan uppträda när kontrollvillkoret utgörs av standardvård och den utvärderade interventionen består av en sammansättning av flera mer eller mindre verksamma komponenter. Detta problem har att göra med att kom-

ponenter, som ingår i den nya och kanske mer effektiva interventionen, börjar spridas och integreras som delar i interventioner som ingår i standardvården (en slags kontaminering). Om detta stämmer borde effekten av ICM i jämförelse med UC bli allt mindre över tid eftersom UC blir allt mer lik ICM över tid. I Figur 9.4 har nya subgrupper bildats där resultaten delats upp i två hälfter i enlighet med medianen (mellan år 1997 och 1998) för det tidsspann som omfattas. Med denna nya indelning försvinner heterogeniteten i båda subgrupperna, skillnaden mellan subgrupperna blir statistiskt signifikant (p=0,003) och andelen av den totala variansen som förklaras av de två subgrupperna blir hela 88,4 procent.



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

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

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

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

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

Dessa två resultat pekar på att kontrollgruppen kan ha kontaminerats över tid. I detta fall när det finns en kontinuerlig variabel som skulle kunna förklara

effektstorlekens variation skulle man även kunna använda sig av metaregression som analysverktyg istället för två tidsperioder [2].

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

Bristande samstämmighet kan inkluderas i metaanalysmodellen

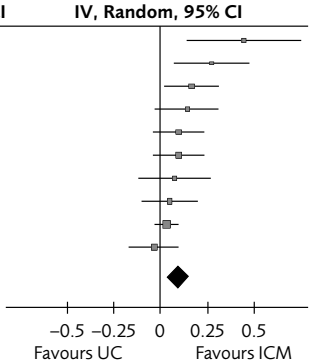
Vi har hittills använt en fixed effect model (FEM) i Figur 9.1–9.4 [2]: under Risk Difference står det Fixed. Denna modell bygger på antagandet att samtliga resultat utgör slumpmässiga urval från en och samma population där det finns en enda ”sann” effekt. Ett annat ganska vanligt sätt att hantera heterogenitet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

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

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

Study or subgroup	Intensive case management Events	Total	Usual care Events	Total	Weight	Risk Difference IV, Random, 95% CI
Lipton 1988	16	20	5	14	3.4%	0.44 (0.14, 0.75)
Bond 1990	30	45	17	43	6.8%	0.27 (0.07, 0.47)
Morse 1992	33	37	48	66	10.3%	0.16 (0.02, 0.31)
Lehman 1997	37	64	27	62	8.4%	0.14 (–0.03, 0.32)
Cox 1998	43	108	25	83	11.3%	0.10 (–0.04, 0.23)
Sosin 1995	49	70	73	121	11.0%	0.10 (–0.04, 0.23)
Morse 2006	25	54	19	49	7.3%	0.08 (–0.12, 0.27)
Rosenheck 2003	41	62	69	113	10.2%	0.05 (–0.10, 0.20)
Conrad 1998	162	178	158	180	19.8%	0.03 (–0.03, 0.10)
Clarke 2000	89	114	40	49	11.6%	–0.04 (–0.17, 0.10)
Total (95% CI)		**752**		**780**	**100.0%**	**0.10 (0.04; 0.16)**
Total events	525		481			

Heterogeneity: Tau²=0.00; Chi²=15.94, df=9 (p=0.07); I²=44%
Test for overall effect: Z=3.11 (p=0.002)



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

Att dela upp resultaten i subgrupper (Figur 9.3 och 9.4) eller inkludera heterogeniteten i metaanalysmodellen (Figur 9.5) är olika sätt att hantera bristande samstämmighet. Vad dessa alternativa strategier innebär blir tydligt när man tolkar resultaten. Resultaten i Figur 9.5, en effekt på 10 procentenheter i riskskillnad, bygger på antagandena att det finns tio olika populationer med vardera en egen sann effekt för varje studie. De tio skilda resultaten antas utgöra slumpmässiga urval av studier från dessa respektive populationer. Den sammanvägda effekten på 10 procentenheter är därför inte en skattning av en sann effekt utan en skattning av medelvärdet i en fördelning av skattade ”sanna” effekter. Uppdelningen i subgrupper (Figur 9.3 och 9.4) istället för att använda REM innebär antaganden om att det finns två populationer, en för vardera subgruppen, och två ”sanna” effekter. Dessa populationer bedöms vara för olika för att det ska vara meningsfullt att inkludera resultat från dem i samma sammanvägning.

Stor klinisk heterogenitet och ingen statistisk sammanvägning

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

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

Analysverktyg eller del i evidensprofil

Metaanalys kan användas på olika sätt. I exemplen ovan har metaanalysen fungerat som ett analysverktyg med vars hjälp man får en bättre förståelse för de data man arbetar med. När väl slutrapporten skrivs och underlaget ska tabuleras i GRADE och bilda en evidensprofil (Kapitel 10), bör man välja de metaanalyser som ska ingå med omsorg. Detta gäller vilka studieresultat som ska ingå, val

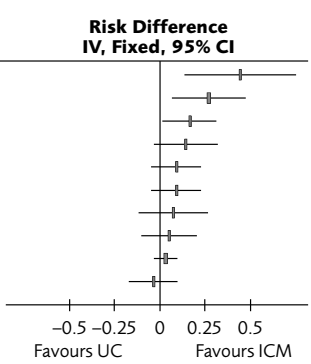
av modell (fixed effects model eller random effects model), eventuella subgrupper samt om man ska göra någon sammanvägning. Även val av statistiska effektmått (oddskvot, riskkvot, riskskillnad, hasardkvot m m) behöver motiveras. Dessa mått har olika statistiska egenskaper och användningsområden. I exemplen ovan har riskskillnad valts av pedagogiska skäl eftersom riskskillnad är lätt att förstå intuitivt.

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

Anta att Figur 9.6 används i det slutgiltiga underlaget som en del i en evidensprofil. I detta fall har resultaten bedömts komma från studier vilka är alltför olika för att en sammanvägning ska vara meningsfull. Det centrala i detta fall är vilket eller vilka resultat som är mest relevanta för beslutsfattande inom den svenska praktiken (jämför överförbarhet i GRADE). Alla tio resultat i Figur 9.6 kanske inte är lika relevanta om population, intervention, kontrollalternativ, effektmått och design beaktas i detalj. Kanske Morse och medarbetare är den studie som bäst fångar de alternativ som praktiken står inför, kanske något eller några andra studieresultat är mest relevanta. Vilket eller vilka resultat som då väljs avgör förstås vilket beslut som stöds på samma sätt varför detta val måste motiveras på ett systematiskt och transparent sätt.

Figur 9.6
Forest plot utan sammanvägning.

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



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

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

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

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

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

Det kan verka poänglöst att genomföra en metaanalys med endast två studier som kanske är för olika för att vägas samman. Som analysverktyg kan metaanalysen ändå ha sin roll. Skillnaden mellan de två studieresultaten blir tydliga. Detta kan göra att man blir uppmärksam på kliniska och metodologiska skillnader man kanske inte uppmärksammat tidigare. Slutligen blir studie-

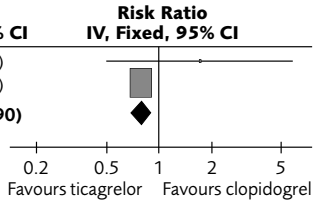
[3] I ticagrelorstudien redovisas effekten som hasardkvot vilket är ett bättre alternativ än riskkvot eftersom tiden till händelsen beaktas. Eftersom dessa uppgifter inte finns i den mindre har vi använt riskkvot.

resultatens relativa statistiska vikt tydlig. Allt detta kan vara till hjälp när man slutligen bedömer vad som ska ingå i evidensprofilen. Om de två studierna bedöms vara tillräckligt lika bör en sammanvägning ingå i evidensprofilen för att förenkla presentation och tolkning (se Kapitel 10 om GRADE).

Figur 9.7
Ticagrelor mot clopidogrel.

Study or subgroup	Ticagrelor Events	Total	Clopidogrel Events	Total	Weight	Risk Ratio IV, Fixed, 95% CI
Cannon DISPERSE2 2007	7	334	4	327	1.1%	1.71 (0.51; 5.80)
Wallentin PLATO 2009	399	9 333	506	9 291	98.9%	0.78 (0.69; 0.89)
Total (95% CI)		**9 667**		**9 618**	**100.0%**	**0.79 (0.70; 0.90)**
Total events	406		510			

Heterogeneity: Chi²=1.56, df=1 (p=0.21); I²=36%
Test for overall effect: Z=3.59 (p=0.0003)



CI = Confidence interval

Metaanalys av observationsstudier

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

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

Om syftet med studien är att utvärdera en intervention i jämförelse med ett kontrollalternativ kan det vara möjligt att använda studieresultaten i en metaanalys. Om huvudsyftet inte varit en sådan utvärdering utan istället att testa

en kausal modell kan det vara svårare att använda resultaten i en metaanalys, speciellt om det inte finns tillräckligt med statistisk information (t.ex. antal individer, medelvärden, spridningsmått).

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

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

Exempel 9.1
Observationsstudie med skillnader i baslinjedata.

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

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

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

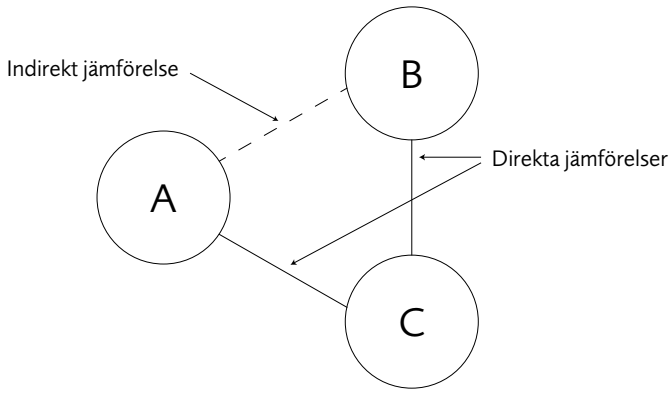
Tabell 9.1
Modell, variabler och matchningsprocedur.

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

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

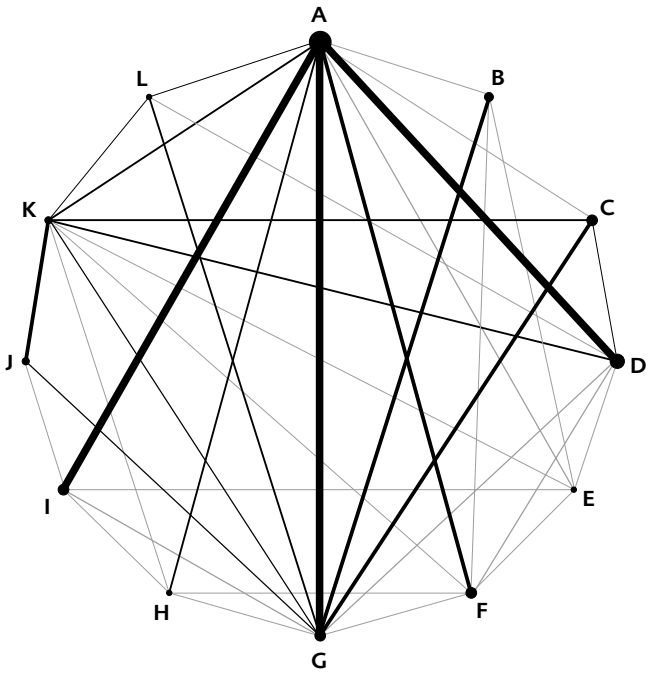
Nätverksmetaanalys

Det har utvecklats metaanalysmetoder för att hantera underlag där direkta jämförelser av relevans för praktiken saknas. Man kan till exempel behöva veta vilka effekterna skulle bli om två interventioner ställdes mot varandra, men där de båda endast jämförts med ett tredje alternativ (Figur 9.8).



Figur 9.8
Exempel på en enkel indirekt jämförelse.

Ofta är indirekta jämförelser mer komplicerade än den i figur 1 och kan bestå av omfattande nätverk med ett stort antal jämförelser (Figur 9.9).



Figur 9.9
Exempel på ett relativt stort nätverk. Större nätverk ökar inferensen då underlaget blir större. Figuren visar ett nätverk av direkta jämförelser där bokstäverna representerar jämförelse-objekt, till exempel olika mediciner, och linjerna representerar jämförelser, där tjockare linjer indikerar fler jämförande studier. Indirekta jämförelser kan beräknas där inga linjer finns.

För att nätverksmetaanalysen ska ge bra resultat krävs:

1. De ingående jämförelserna som till exempel patientegenskaper och andra effektmodifierare ska vara likartade(likhetsantagande).
2. Det får inte finnas relevant heterogenitet avseende studieresultaten (homogenitetsantagande).
3. Det får inte finnas relevanta motsägelser mellan effekterna i de direkta och indirekta jämförelserna (jämförbarhetsantagande) [36].

De första två antagandena gäller även för traditionell metaanalys.

Sedan cirka 2009 är den vanligaste metoden vid nätverksmetaanalys baserad på bayesianska hierarkiska modeller. Tidigare var Buchers metod och meta-regression[4] vanligare.

Vid nätverksmetaanalys görs flera jämförelser parallellt. Som i exemplet i Figur 9.8 där intervention A jämförs med B, B med C samt A med C. Här är endast jämförelsen mellan A och B indirekt men flera jämförelser kan vara indirekta så länge det finns tillräckligt många direkta jämförelser som stödjer beräkningen. Dessutom kan jämförelserna vara direkta och indirekta samtidigt, så kallade mixade jämförelser. Ett grundläggande antagande vid nätverksmeta-analys är att direkta och indirekta jämförelser skattar samma parameter, det vill säga en direkt mätning av den relativa effekten mellan A och B i exemplet är densamma som den indirekta mätningen av den relativa effekten mellan A och B. Den indirekta jämförelsen mellan A och B estimeras med direkta mätningar mellan A och C samt C och B.

Nätverksmetaanalystekniker stärker inferensen på den relativa effekten av två behandlingar genom att inkludera direkta och indirekta jämförelser och samtidigt tillåta inferens på samtliga behandlingar med bibehållen randomisering [37].

Metaanalys av diagnostiska studier

Ytterligare ett tillämpningsområde för metaanalyser är *diagnostik* [2].

Det viktigaste utfallsmåttet för studier av diagnostiska metoder är effekter på deltagarnas hälsa. Sådana studier bör vara randomiserade och gängse metodik för metaanalyser kan användas. Ofta saknas studier som utvärderar effekter av att använda ett test utan man får nöja sig med att utvärdera testets diagnostiska tillförlitlighet (Se Kapitel 7).

Metaanalys av studier om diagnostisk tillförlitlighet är mera komplicerat än för interventionsstudier. Orsaken är att det finns två utfallsmått, sensitivitet och specificitet, som är korrelerade till varandra (Kapitel 7). En tidigare använd metod var kopplade forest plots, dvs diagrammet visade metaanalyser för sensitivitet respektive specificitet parallellt (Se Kapitel 7, Figur 7.3). Metoden förutsatte att samtliga studier använde samma tröskelvärde.

Kopplade forest plots kan vara värdefulla som ett första steg för att få en uppfattning om hur heterogent materialet är. Det finns möjligheter att göra kopplade forest plots i Cochrane collaborations programvara RevMan men observera att diagrammet inte kommer att ge sammanvägda värden för sensitivitet och specificitet. Om den kopplade forest plot visar hög heterogenitet mellan studierna kan det finnas skäl att analysera orsaker till heterogeniteten och överväga om alla studier ska ingå.

[4] Endast lämplig för random-effects models enligt Higgins and Thompson, 2004, Statistics in medicine.

Nackdelen med kopplade forest plots är att de inte tar hänsyn till det inbördes beroendet mellan sensitivitet och specificitet, tröskeleffekten. Resultaten kan istället vägas samman med hjälp av antingen bivariat metaanalys eller Hierarkisk summa ROC-analys (HSROC). De beskrivs utförligt i Cochranes handbok, Kapitel 10 [24]. Metaanalyserna kan inte genomföras helt inom programvaran RevMan utan kräver också tillgång till statistisk programvara av typen Stata, SAS eller SPSS (senare versioner).

Programvaror

Det finns flera olika program som kan användas för metaanalys. Ett av de mest lättanvända programmen, som för närvarande är fritt tillgängligt på internet, är Review Manager (RevMan) som har tagits fram inom Cochrane Collaboration. Programmet följer internationellt etablerade konventioner men klarar inte mer komplicerade former av metaanalyser som till exempel diagnostiska metaanalyser eller nätverksmetaanalys.

Något mer avancerad är Comprehensive Meta-Analysis (CMA) som har fler funktioner än RevMan (t.ex. metaregression), men är avgiftsbelagd.

Meta-DiSc som utvecklats speciellt för metaanalyser inom diagnostik är ännu fritt tillgängligt via internet.

Ytterligare ett par fritt tillgängliga programvaror för metanalys, men som dessutom har ett brett register av tillämpningar inom statistisk analys i övrigt är R tillsammans med ett stort urval av programpaket, samt Python tillsamans med programarkiven NumPy, SciPy och PANDAS.

Även STATA och SAS är kompetenta verktyg för statistisk analys bland annat metaanalys. Dessa två är dock inte fritt tillgängliga.

Några programvaror som är lämliga för nätverksmetaanalys är WinBUGS, OpenBUGS, JAGS, Python tillsammans med PyMc, Stan, JULIA tillsammans med Mamba, R tillsammans med WinBUGS, OpenBUGS, JAGS, Laplaces demon, STATA och SAS. Samtliga dessa programmeringsverktyg, utom STATA och SAS, är fritt tillgängliga på internet.

Ett område i snabb utveckling

Metaanalyser och relaterade metoder är föremål för en snabb utveckling. Gamla arbetssätt modifieras och nya metoder tillkommer.. I dessa sammanhang är det av stor vikt att följa utvecklingen via internationella nätverk såsom Cochrane Collaboration och GRADE Working Group, där konventioner, systematik och transparens utvecklas. Av speciell betydelse för arbetet med metaanalyser är PRISMA statement (en vidareutveckling av QUORUM statement). Tre förändringar det senaste decenniet kan lyftas fram: (a) fokus har förskjutits från

enskilda studier till effektmått (vilka kan inkludera resultat från flera studier) då risk för bias bedöms, (b) betydelsen av kontext och extern validitet har betonats mer än tidigare, och (c) gamla former av evidenshierarkier har börjat problematiseras (vilket innebär att det inte är omöjligt att resultat från observationsstudier kan bedömas ha låg risk för bias).

Exempel 9.2
Kopplade forest plots av sensitivitet och specificitet av fem studier som undersöker tillförlitligheten hos kyla-test (en pellet doppad i etylklorid appliceras på tandytan) för att bestämma om en tandpulpa är vital eller non-vital.

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

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

Referenser

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

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

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

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

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

9. 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.

10. 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.

11. 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.

12. 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.

13. 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.

14. 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.

15. 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.

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

17. 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 of Systematic Reviews 2009, Issue 1. Art. No.: MR000006. DOI: 10.1002/14651858. MR000006.pub3.

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

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

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

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

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

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

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

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

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

27. Zelada MA, Salinas R, Baztán JJ. Reduction of functional deterioration during hospitalization in an acute

geriatric unit. Arch Gerontol Geriatr 2009;48:35-9.

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

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

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

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

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

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

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

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

36. Kiefer C, Sturtz S, Bender R. Indirect comparisons and network meta-analyses. Dtsch Arztebl Int 2015;112:803-8.

37. Bayesian Hierarchical Methods for Network Meta-Analysis. A dissertation submitted to the faculty of the graduate school of the university of Minnesota.By Jing Zhang. In partial fulfillment of the requirements for the degree of Doctor of Philosophy Advised by Haitao Chu July, 2014.