Att göra effektutvärderingar

Att göra effektutvärderingar

Red. Knut Sundell

GOTHIA
FÖRLAG

ISBN 978-91-7205-838-5

Redaktör: Hanna Håkansson
Omslag: Tomas Rudström
Grafisk form: RPform, Richard Persson

Första upplagan, första tryckningen
Tryck: Scandbook AB, Falun 2012
Tryckt på miljövänligt framställt papper.

Gothia Förlag
Box 22543, 104 22 Stockholm
Kundservice 08-462 26 70, Fax 08-644 46 67
www.gothiaforlag.se

Innehåll

Förord

Årligen berörs hundratusentals personer i Sverige av psykosociala och pedagogiska interventioner (behandlingar, insatser, metoder, åtgärder). Det handlar ibland om interventioner som inte är frivilliga, som ställer stora krav på de medverkande och där målgruppen till en del utgörs av samhällets mest utsatta. Endast i undantagsfall finns en vetenskaplig kunskap om vilka effekter dessa interventioner har. Socialstyrelsen har därför tagit initiativ till denna bok om hur effektutvärderingar kan bedrivas. Boken fokuserar på randomiserade och icke-randomiserade kontrollerade experiment. Andra typer av effektutvärderingar kan också ge värdefull kunskap om hur olika interventioner fungerar men behandlas inte närmare i denna bok. Målgrupp för boken är forskare från alla discipliner som är intresserade av att lära sig mer om effektutvärderingar, både forskarstuderande och forskare som redan ansvarat för effektutvärderingar.

Boken utgör ett samarbetsprojekt mellan Socialstyrelsen och Atferdssenteret i Oslo. Bokens disposition planerades av Kristin Amlund Hagen, Terje Ogden samt undertecknad. Under våren 2011 medverkade också Martin Forster som redaktör, innan han övergick till andra arbetsuppgifter utanför Socialstyrelsen. Bokens 21 författare har lång erfarenhet av effektutvärderingar och representerar olika discipliner som psykologi, folkhälsovetenskap, socialt arbete, pedagogik, statistik, epidemiologi, sociologi, evolutionsbiologi, medicinsk vetenskap och datavetenskap.

Texten har fackgranskats av professor emiritus Bengt-Åke Armelius och professor emerita Kerstin Armelius, institutionen för psykologi vid Umeå universitet; professor emiritus Kjell Hansson, Socialhögskolan vid Lunds universitet; fil.dr. i sociologi Ulla Jergeby samt

professor och överläkare Per-Olof Östergren, Socialmedicin och global hälsa vid Lunds universitet. Kapitel 6 har fackgranskats av Matilda Hansson, Socialstyrelsen, och kapitel 12 av Åke Hellström, professor emeritus vid institutionen för psykologi vid Stockholms universitet. Boken har språkgranskats av Eva Häggmark, med bistånd från Cecilia Andrée Löfholm, Ulrika Bergström, Malin Hultman, Per Arne Håkansson, Catrine Kaunitz och Elizabeth Åhsberg, samtliga från Socialstyrelsen.

Bokens innehåll

Boken består av 19 kapitel. De första två beskriver centrala begrepp för effektutvärderingar samt ger en kortfattad historik över experimentella utvärderingar. Därefter följer 16 kapitel fördelade på fyra delar, som grovt motsvarar den ordning som en effektutvärdering genomförs på (se figur): planering (kapitel 3–7), datainsamling (kapitel 8–10), analyser (kapitel 11–17) samt rapportskrivning (kapitel 18).

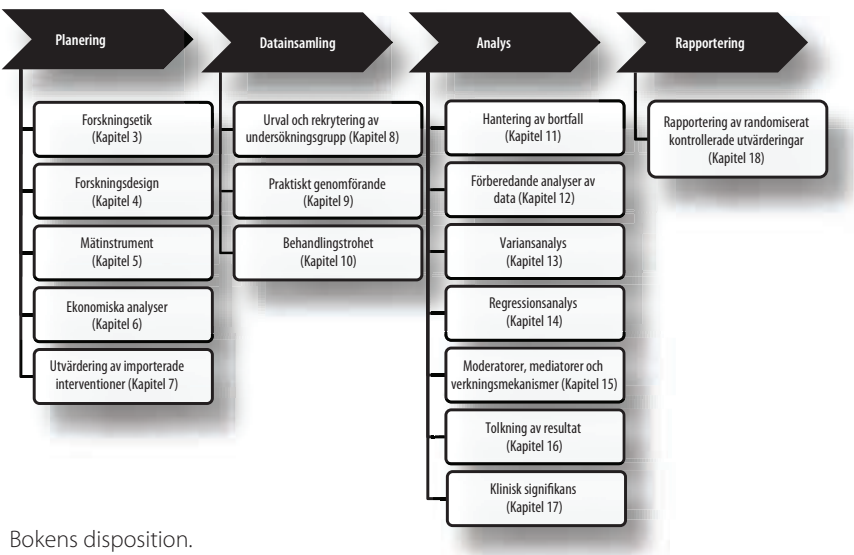
Fem kapitel behandlar förberedelser inför datainsamlingen. Kapitel 3 handlar om etik i utvärderingsarbete – vad kan man göra som forskare och vad är oetiskt?

Kapitel 4 behandlar den centrala frågan om val av undersökningsdesign. Kapitlet inleds med en redogörelse för olika typer av validitet och hot mot dem. Frågor som berörs därefter är val av jämförelsegrupp, antal och tidpunkt för undersökningstillfällen samt beräkning av hur många undersökningspersoner som behöver rekryteras för att få acceptabel statistisk styrka.

Kapitel 5 handlar om val av mätmetoder för att avläsa förändring. En utvärdering är inte starkare än de effektmått som används.

De följande två kapitlen behandlar effektutvärderingar ur två ofta förbisedda perspektiv. Kapitel 6 behandlar kostnadsanalyser. En intervention kan vara mindre effektiv än ett alternativ men ändå vara mer attraktiv, eftersom den är kostnadseffektiv. Kostnadsanalyser av hög kvalitet förutsätter i allmänhet prospektiv datainsamling.

I kapitel 7 behandlas strategier för att utvärdera ”importerade” interventioner som identifierats som effektiva i andra kontexter. Dessa interventioner kan behöva anpassas till en ny kontext innan de kan implementeras. Kapitlet beskriver hur en effektutvärdering kan planeras för att bättre kunna analysera om effekterna beror på interventionen, anpassningen av interventionen till en ny kontext, behandlingstrohet eller på andra faktorer som kulturella och sociala skillnader.



Bokens disposition.

Nästa tre kapitel berör datainsamling. Kapitel 8 berör rekrytering av undersökningsgrupp. Ett vanligt problem i effektutvärderingar är att få en tillräckligt stor undersökningsgrupp. Kapitlet diskuterar vanliga svårigheter och vad som kan underlätta rekryteringen av studiedeltagare.

Kapitel 9 handlar om det praktiska genomförandet av datainsamlingen, inklusive hur insamlad information sammanställs och förvaras. Detta område har sällan eller aldrig behandlats systematiskt för utvärderingsforskning.

Kapitel 10 beskriver vikten av att under studiens gång registrera

om interventionerna som undersöks har implementerats på rätt sätt. Om inte undersökningspersonerna får den intervention de skulle ha fått, kan felaktiga slutsatser dras om interventionens effekter. Detta gäller även kontrollbetingelsen: Har rätt intervention getts eller har jämförelsegruppens personer kanske fått samma intervention som de i interventionsgruppen?

De följande sju kapitlen 11 till och med 17 behandlar analysfasen. Kapitel 11 handlar om problemet att personer som medverkar i effektutvärderingar ibland i förtid avbryter deltagandet, alternativt inte besvarar alla frågor. Kapitlet beskriver olika sätt att hantera detta bortfall så att det inte snedvrider resultaten.

Kapitel 12 handlar om kontroll av datas kvalitet, bland annat identifikation av extremvärden, variablers samvariation samt om data är normalfördelade.

Kapitel 13 beskriver de vanligaste formerna av variansanalys och kovariansanalys.

Kapitel 14 behandlar olika typer av regressionsanalyser. Dessa analyser är särskilt lämpade när undersökningspersonerna är beroende av varandra, exempelvis elever som går i samma klass och klienter som får behandling av samma terapeut. Om dessa beroenden inte hanteras korrekt kan det snedvrida tolkningen av resultaten.

Kapitel 15 handlar om analyser av om en intervention är mer effektiv för en viss grupp personer än för andra (moderatorer) och om test av en interventions kärnfaktorer (mediatorer). Det sista är speciellt viktigt i händelse av att en intervention inte visat sig ge starkare effekter än sin kontroll.

Kapitel 16 berör olika statistiska mått för att redovisa resultat, som statistisk signifikans, effektstorlek och oddskvot.

I kapitel 17 behandlas klinisk signifikans. En intervention kan vara statistiskt mer effektiv än sin kontroll men utan att vara kliniskt meningsfull.

Kapitel 18 behandlar hur en experimentell effektutvärdering ska avrapporteras vetenskapligt. Utgångspunkten är de så kallade CONSORT statements. Utan en korrekt avrapportering finns en

risk för felaktiga slutsatser. Rubin och Parrish[1] har exempelvis visat att i 70 procent av alla vetenskapliga artiklar i två tidskrifter för socialt arbete drog författarna kausala slutsatser som den använda vetenskapliga designen inte medgav. Detta kapitel sammanfattar mycket av bokens övriga innehåll.

Kapitel 19 presenterar tankar om effektutvärderingar i framtiden.

Boken avslutas med en ordlista över termer som rör effektutvärderingar.

Eftersom bokens författare representerar olika discipliner varierar språkbruket något mellan kapitlen. Exempelvis används klient, patient, brukare och person för att beteckna dem som blir föremål för interventioner. På samma sätt används professionella, behandlare, praktiker och terapeut som term för dem som ansvarar för interventionen. De som genomför en studie benämns forskare, projektledare och experimentledare. Ambitionen är att varje kapitel ska kunna läsas fristående från de andra, och därför kan det förekomma vissa upprepningar.

Stockholm, 2012
Knut Sundell, Socialstyrelsen

1 Rubin, A. & Parrish, D. (2007). Problematic phrases in the conclusion of published outcome studies: Implications for evidence-based practice. *Research on social work practice, 17*, 334–347.

Medverkande författare

STEN ANTTILA, filosofie doktor i sociologi, arbetar som projektledare vid Statens beredning för medicinsk utvärdering (SBU). Hans huvudområde är systematiska översikter och metaanalys. Han var tidigare verksam vid Uppsala universitet, Örebro universitet, Mittuniversitetet, Ersta Sköndal högskola, Institutet för utveckling av metoder i socialt arbete (IMS) samt Socialstyrelsen.

GUNNAR BJØRNEBEKK, filosofie doktor i motivationspsykologi från Universitetet i Oslo. Han har sedan 2007 arbetat som forskare på Atferdssenteret med utvärderingar av evidensbaserade program mot allvarliga beteendeproblem bland barn och ungdom. Han arbetar nu med validering av mätinstrument för olika typer av aggression, en undersökning av personlighetsdrag och typ av kriminalitet hos medlemmar i antisociala gäng samt en effektutvärdering av Parent Management Training, Oregonmodellen (PMTO) för barn med adhd.

MARIA BODIN, filosofie doktor i folkhälsovetenskap och legitimerad psykolog. Mellan 2007 och 2010 ansvarade hon för tre effektutvärderingar av preventionsprogram för barn och unga på STAD (Stockholm förebygger alkohol- och drogproblem), Centrum för Psykiatriforskning Stockholm. Sedan 2011 är hon anställd som forskare vid enheten för kunskapsutveckling vid Socialstyrelsen i Stockholm.

MARIA PAOLA CARIA, biostatistiker och legitimerad tandläkare, är knuten till universitetet i Novara i Italien. Hon är också doktorand vid Institutionen för folkhälsovetenskap, Karolinska Institutet. Hon har genomfört flera analyser av epidemiologiska data och klusterbaserade effektutvärderingar av prevention i Europa.

BASSAM MICHEL EL-KHOURI, filosofie doktor i psykologi samt filosofie licentiat i data- och systemvetenskap. Han är anställd som forskare på enheten för kunskapsöversikter vid Socialstyrelsen i Stockholm. Tidigare har han forskat vid Stockholms universitet (psykologi), Kungliga Tekniska högskolan (data- och systemvetenskap) och Karolinska Institutet (folkhälsovetenskap) samt FoU-enheten i Stockholms stad. Hans arbetsområde är systematiska översikter med närliggande metod- och statistikfrågor. Hans forskningsintressen inkluderar framtagande av personorienterade metoder för att studera longitudinell utveckling och tillämpningen av dessa i olika domäner.

LAURA FERRER-WREDER, filosofie doktor i psykologi och docent vid institutionen för psykologi vid Stockholms universitet. Hon har tidigare varit anställd vid Pennsylvania State University, Capital College (2001–2005) och Barry University i USA (2005–2009). Hennes forskning handlar om att utveckla förebyggande interventioner riktade till barn och ungdomar. Hon har medverkat i 13 effektutvärderingar. Hennes nuvarande forskning är inriktad på att främja positiv ungdomsutveckling genom samarbete mellan universitet och samhälle.

MARTIN FORSTER är legitimerad psykolog och filosofie doktor i psykologi. Han har forskat om effekter av förebyggande interventioner via föräldrar och lärare för barn med utagerande problematik samt genomfört ett antal randomiserade studier inom området. Han har också varit ansvarig för utveckling av flera manualbaserade program (Komet för lärare, Komet för föräldrar till barn 3–11 år, Komet för föräldrar till barn 12–18 år, Förstärkt Komet, InternetKomet och Familjeverkstan). I samband med det har han varit engagerad i frågor som rör behandlingstrohet.

MARIA ROSARIA GALANTI, medicine doktor och filosofie doktor i cancerepidemiologi, är docent i epidemiologi vid institutionen för folk-

hälsovetenskap, Karolinska Institutet. Hon har varit huvudansvarig forskare i två stora skolbaserade effektutvärderingnar för att studera resultatet av förebyggande interventioner. Hon är ledamot i styrgruppen för *European Society for Prevention Research* (www.euspr.org).

JAMES JACCARD är professor i psykologi vid Florida International University. Han har tidigare arbetat vid Purdue University (1976–1981) och the State University of New York, Albany (1982–2004). Han är författare till mer än 100 vetenskapliga artiklar, bland annat om ungas problembeteenden och hälso- och sjukvård. Han har också skrivit flera böcker om statistik. Hans senaste bok som publicerades 2010 tillsammans med Jacob Jacoby heter *Theory construction and model-building skills: A practical guide for social scientists*.

HARALD JANSON är forskare vid Atferdssenteret, docent i psykologi och privatpraktiserande psykolog. Tillsammans med andra forskare vid Atferdssenteret genomför han det longitudinella forskningsprojektet *Barns sosiale utvikling,* som följer 1 159 barn i fem norska kommuner från sex månaders ålder och uppåt med avseende på social utveckling. Hans forskningsintressen är psykometri, personlighetsmätning och longitudinella studier.

BJØRN ARILD KRISTIANSEN har magisterexamen i datavetenskap och är datachef på Atferdssenteret i Oslo. Han är ansvarig för datahantering och stöd till forskning vid Atferdssenterets forskningsavdelning. Han har bland annat utvecklat kontaktdatabaser och system för elektronisk datainsamling för flera av Atferdssenterets forskningsprojekt. Han ingår i Atferdssenterets logistikteam.

CHRISTINE HASSEL KRISTOFFERSEN är cand. polit. i psykologi och forskningskonsult vid Atferdssenteret. Hon har varit projektledare och medansvarig för planering, genomförande och drift av den longitudinella studien *Barns sosiale utvikling*. Hon har också utvecklat allmänna riktlinjer och rutiner för planering och ledning av forsk-

ningsprojekt vid Atferdssenteret. Christine ingår i Atferdssenterets logistikteam.

KRISTIN BERG NORDAHL är chef för Atferdssenterets logistikteam och doktorand i utvecklingspsykologi vid universitetet i Bergen. Hon har ansvarat för överföring och anpassning av observationer i samverkan med Atferdssenterets utvärdering av implementering och effekter av Parent Management Training, Oregonmodellen (PMTO) i Norge. Under senare år har hon arbetat med systematiska observationer av samspelet mellan föräldrar och barn i den longitudinella studien *Barns sosiale utvikling*.

TERJE OGDEN är forskningsdirektör vid Atferdssenteret, Unirand, och professor vid Psykologiska Institutionen vid Universitetet i Oslo. Han har sin filosofiska doktorsgrad från Universitetet i Bergen och har under flera år arbetat med implementering och forskningsbaserad utvärdering av evidensbaserade program inriktade mot allvarliga beteendeproblem bland barn och ungdom.

ASGEIR RØYRHUS OLSETH är forskningskonsult vid Atferdssenteret och masterstudent i pedagogisk forskningsmetodik vid Universitetet i Oslo. Han är projektkoordinator och medansvarig för planering, genomförande och drift av studien *Positiv atferdsstøtte i skolen*. Han har även arbetat med datainsamling i Atferdssenterets effektutvärdering av Parent Management Training, Oregonmodellen (PMTO). Han ingår i Atferdssenterets logistikteam.

TINA OLSSON är filosofie doktor i socialt arbete och Master of Public Administration. Hon har i 20 års tid arbetat med frågor om barn, unga och familjer som socialarbetare och utvärderare i USA och Sverige. Hon har medverkat i ett antal effektutvärderingar, och i dag fokuserar hennes forskning på ekonomiska utvärderingar.

ANN-CHARLOTTE SMEDLER är legitimerad psykolog och professor i psykologi vid Stockholms universitet. Hennes forskning rör utveckling och psykisk hälsa hos barn och unga som utsatts för särskilda risker under uppväxten. Hon har medverkat i flera utredningar av metoder för att behandla och förebygga psykiska problem hos barn och är ledamot i Regionala etikprövningsnämnden i Stockholm och i Vetenskapsrådets expertgrupp för etik.

SANDRA SOUTH är filosofie doktor i evolutionsbiologi och forskare vid enheten för kunskapsutveckling vid Socialstyrelsen. Hon har medverkat i forskningsprojekt vid Uppsala universitet, University of North Carolina, University of Exeter, University of Melbourne, Københavns Universitet samt University of California. Hon har genomfört ett stort antal experiment för att utforska djurs sexuella beteende ur ett evolutionsperspektiv. Som lärarassistent i statistik för naturvetare vid Uppsala universitet har hon hjälpt över hundra forskarstudenter att tillämpa variansanalys.

KNUT SUNDELL, docent i psykologi, socialråd och chef för enheten för kunskapsutveckling vid Socialstyrelsen. Han var forskningsledare på Stockholms stads Forsknings- och utvecklingsenhet mellan 1990 och 2006 och chef för Institutet för utveckling av metoder i socialt arbete (IMS) mellan 2007 och 2009. Han har ansvarat för ett 15-tal effektutvärderingar inom främst barn- och ungdomsområdet.

RIKARD WICKSELL, doktor i medicinsk vetenskap och legitimerad psykolog, har sedan 2001 arbetat vid Karolinska Universitetssjukhuset med att utveckla en beteendemedicinsk behandlingsmodell baserad på Acceptance and Commitment Therapy för patienter med långvarig, handikappande smärta. Han arbetar i dag med klinisk forskning rörande effekter och förändringsprocesser vid beteendemedicinsk behandling.

LARS-GÖRAN ÖST, legitimerad psykolog, legitimerad psykoterapeut, handledare och professor i klinisk psykologi vid Stockholms universitet. Han har de senaste 40 åren arbetat med psykoterapiforskning inom schizofreni, narkomani, övervikt, stamning, specifika fobier, social fobi, agorafobi, panikstörning, generaliserat ångestsyndrom, posttraumatiskt stressyndrom för vuxna samt specifika fobier och social fobi hos barn.

Kapitel 1 | Knut Sundell & Terje Ogden

Introduktion

> Det sker många fler experiment i vår vardag än i världens alla forskningslaboratorier, men det mesta av detta experimenterande genomförs i det tysta och utan att öka vår kunskap.

Ovanstående citat är hämtat från en bok från 1932 av Sidney och Beatrice Webb (Oakley, 1998) och syftar på att nya arbetssätt i allmänhet introduceras utan vetenskapligt stöd och utan försök att undersöka vilka effekter de har för en grupp individer.

Den här boken handlar om metoder för att skapa kunskap om interventioners (behandlingsmetoders, insatsers) effekter. För det räcker inte med vällovliga teorier, goda avsikter, hårt arbete, enighet om att en viss intervention är effektiv, att det känns rätt, att man alltid har gjort på det sättet eller att någon auktoritet säger att det är bra (Chaffin & Friedrich, 2004). Redan 1957 konstaterade Julian C. Stanley: ”Experters utlåtanden, samlade bedömningar, brillianta insikter och skarpsinniga antaganden är ofta vilseledande.”

Påståenden om en interventions effekt bör alltid bemötas med frågan: Hur vet man det? Det vetenskapligt starkaste sättet att svara på frågan är genom en experimentell studie som jämför en grupp personer som får en intervention med en annan grupp som antingen inte får någon intervention alls eller som får en annan intervention. Om studien är randomiserad, så att slumpen bestämmer vem som hamnar i respektive grupp, ökar förutsättningarna för att avgöra in-

terventioners verkliga effekter. Randomiseringen gör i teorin att alla andra faktorer som påverkar människors liv blir lika representerade i båda grupperna, vilket lämnar interventionen som den enda systematiska skillnaden mellan grupperna. Även icke-randomiserade effektutvärderingar kan ge trovärdiga resultat.

Inom framför allt medicin har experimentella effektutvärderingar en stark legitimitet. En viktig orsak till det är insikten om att läkares beslut måste regleras för att säkra bästa vård för patienterna. För de flesta av oss är det otänkbart att valet av behandling skulle bestämmas av läkarens personliga preferenser.[2] Inom den sociala sektorn och i skolan förekommer det dock att professionella använder interventioner som de själva föredrar och oberoende av om det finns ett vetenskapligt stöd för dem eller om de passar klientens eller elevernas behov.

Effektutvärderingars främsta uppgift är inte att visa att något fungerar fullt ut utan att med rimlig säkerhet bevisa att det fungerar och därmed erbjuda allmänheten skydd mot ett skadligt experimenterande med människors liv. Det finns flera exempel på psykosociala interventioner som orsakat skada, till och med tagit människors liv (t.ex. Beutler, 2000; Kennedy, Mercer, Mohr & Huffine, 2002; Lilienfeld, 2007; Sundell & Vinnerljung, 2008). Ett vanligare problem är att många interventioner saknar påvisbara effekter. Ett exempel är att 63 procent av 46 interventioner för unga kriminella saknade belägg för att göra skillnad (Weisburd, Lum & Petrosino, 2001).

Metodiken att göra effektutvärderingar har utvecklats påtagligt under senare år. För att utvärderingsresultat ska få genomslag i praktiken krävs att utvärderingar håller hög kvalitet. Detsamma gäller för att kunna publicera sig i välrenommerade vetenskapliga tidskrifter. Syftet med den här boken är att ge praktiskt stöd i de olika moment som ingår i en effektutvärdering av hög kvalitet.

I detta introducerande kapitel behandlas viktiga begrepp med anknytning till effektutvärderingar.

2 För avskräckande exempel, se Singh & Ernst, 2008.

Förekomst av effektutvärderingar

Antalet personer i Sverige som årligen berörs av psykosociala och pedagogiska interventioner är flera hundra tusen. I de allra flesta fall saknas vetenskaplig kunskap om dessa interventioner. I USA har det exempelvis beräknats att endast fem procent av alla ungdomsbrottslingar blir föremål för interventioner som effektutvärderingar har visat är effektiva (Greenwood, 2008).

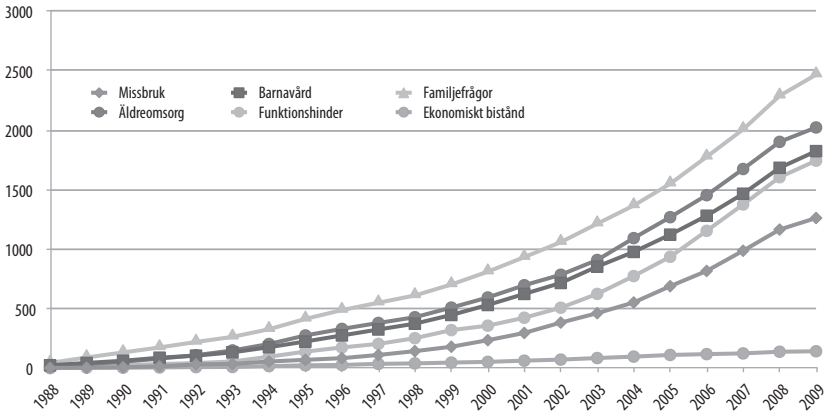
Sommaren 2010 fanns det endast 51 svenska effektutvärderingar av sociala interventioner publicerade i vetenskapliga tidskrifter (Socialstyrelsen, 2011a). Dessa utvärderingar behandlar ett fåtal av alla interventioner som används i Sverige. Socialstyrelsen (2009a) har exempelvis identifierat 103 socialtjänstbaserade interventioner inom öppenvården för barns psykiska hälsa. Av dem är tio utvärderade i Sverige på ett sådant sätt att det går att bedöma deras effekter. Situationen är liknande inom förskola och grundskola; nästan ingen av de pedagogiska metoder som används för barns psykiska hälsa har något säkert vetenskapligt stöd (Socialstyrelsen, 2009b; 2009c).

Antalet personer som årligen får interventioner i Sverige

- 800 000 personer får interventioner inom ramen för socialtjänsten.
- 630 000 personer får interventioner inom psykiatrin.
- 860 000 barn möter pedagogiska interventioner i förskolan.
- 890 000 elever möter pedagogiska interventioner i grundskolan.

Få av dessa interventioner har utvärderats vad gäller effekter.

Det låga antalet svenska effektutvärderingar är inte överraskande när man betänker att få svenska forskare tränats i denna typ av forskning. En genomgång av tio års (1997–2006) avhandlingar från sju vetenskapliga discipliner (folkhälsovetenskap, kriminologi, omvårdnadsvetenskap, pedagogik, psykologi, socialt arbete och sociologi) visar att endast fyra procent av 1 402 avhandlingar inkluderade studier av rimlig kvalitet som utvärderade interventioners effekter (Sundell & Stensson, 2010).



Figur 1:1. Kumulativt antal randomiserat kontrollerade effektutvärderingar (träffar) (Socialstyrelsen, 2011a).

Internationellt sett har antalet effektutvärderingar inom det psykosociala området ökat dramatiskt. En sökning efter randomiserat kontrollerade utvärderingar (RCT) mellan 1988 och 2009 i de viktigaste referensdatabaserna för socialt arbete (PsycINFO, PubMed och Sociological Abstracts/Social Services Abstracts) visar en tydlig ökning av antalet träffar för fem av sex granskade områden (efter justeringar för överlappningar i olika databaser). Det handlar om en årlig ökning om ungefär 15 procent (figur 1:1). Det enda område för vilket antalet träffar endast ökat svagt är ekonomiskt bistånd.

Effektutvärdering

Effektutvärderingar (eng. *outcome research, impact evaluation* och *intervention research, effectiveness trials*) handlar dels om att avgöra om någon förändring sker i människors situation, dels att avgöra om en specifik intervention orsakat det som hänt.

Inom hälso- och sjukvård och folkhälsa beskrivs effektutvärderingar som forskning om interventioners effekter, som beskriver slutresultatet av dessa interventioner, som fokuserar på de berörda

personernas förbättrade hälsa, minskad sjuklighet eller dödlighet och förbättring av onormala tillstånd (till exempel förhöjt blodtryck) och som är avsedda att användas som vetenskapligt underlag för beslut som fattas av dem som deltar i vården (t.ex. Jefford, Stockler & Tatersall, 2003). Anpassat till det samhällsvetenskapliga området kan man säga att effektutvärderingar fokuserar på de berörda personernas förbättrade hälsa och psykosociala situation. För det utbildningsvetenskapliga området handlar det om att främja inlärning och förebygga psykosociala problem genom individuella och gruppinriktade interventioner.

Effektutvärderingar

- studerar interventioners effekter
- fokuserar på de berörda personernas förbättrade hälsa och psykosociala situation
- beskriver slutresultatet av dessa interventioner
- är avsedda att användas som underlag för beslut som fattas av dem som deltar i vård, omsorg och pedagogisk verksamhet.

Viktiga frågeställningar är (1) om behandling är bättre eller sämre än ingen behandling, (2) om en viss intervention är bättre eller sämre än en annan, (3) om mängden intervention är relaterad till effekterna, (4) om effekten är stabil eller förändras över tid samt om (5) effekterna av en intervention motiverar kostnaderna. Andra viktiga frågeställningar är (6) om interventionen använts som tänkt, (7) om interventionen är lika effektiv för olika målgrupper samt (8) om den tänkta mediatorn, förändringsfaktorn, är relaterad till utfallet.

Genom att koppla den vård och omsorg människor får till effekter har effektutvärderingar en nyckelroll i att utveckla, övervaka och förbättra kvaliteten i vård, omsorg och pedagogisk verksamhet. Effektutvärderingar syftar till att ge kunskap om vilka interventioner som fungerar bäst för varje klient (patient och brukare) och under vilka omständigheter. Denna kunskap är viktig eftersom den tillåter professionella att fatta välgrundade beslut baserade på vetenskapliga fakta och ger professionella bättre förutsättningar att ge

rätt behandling till rätt klient, vid rätt tillfälle och med hänsyn till klientens individuella egenskaper.

Uppföljning och utvecklingsarbete

Utöver effektutvärderingar finns det andra aktiviteter som undersöker interventioners värde. En sådan är uppföljning (eng. *monitoring*) som handlar om att fortlöpande samla in information för att kontrollera hur en verksamhet genomförs. Underlaget kan vara information från administrativa register, exempelvis andelen barn och unga som varit föremål för social barnavård under ett år. Uppföljningar görs normalt av tjänstemän inom ramen för en ordinarie verksamhet för att kontrollera att verksamheten uppfyller sitt syfte. En skillnad mellan uppföljning och utvärdering är att uppföljning sker kontinuerligt medan utvärdering är en tillfällig aktivitet. Datainsamlingen vid utvärdering är också i allmänhet mer djuplodande och baseras ofta på uppgifter som inte rutinmässigt samlas in av den verksamhet som utvärderas, till exempel olika former av standardiserade frågeformulär.

En annan närliggande aktivitet är utvecklingsarbete, som kan ske med eller utan utvärdering. Evidensbaserade metoder bygger exempelvis på ett systematiskt utvecklingsarbete, där utvecklingsarbetet utvärderas och erfarenheterna från utvärderingen styr fortsatt utvecklingsarbete som i sin tur utvärderas. Det finns exempel där utvärdering blivit en så integrerad del av utvecklingsarbetet att det är svårt att skilja de två åt. Trots att utvärdering och utvecklingsarbete ibland är nära förknippande med varandra bör de särskiljas, eftersom de har olika mål. Utvärderingens mål är att värdera effekterna av en intervention medan utvecklingsarbetets mål är att genom en serie sammankopplade utvärderingar skapa en effektiv intervention. Initialt kan det handla om små single-case-studier och studier av om interventionen är acceptabel för dem som berörs. Därefter kan pre-post-studier vara relevanta för att försäkra sig om att effekterna går i rätt riktning (d.v.s. inte är skadliga). Nästa steg kan vara att testa interventionen i en eller flera modellutvärderingar (eng. *efficacy*)

och slutligen utvärdering i ordinarie verksamhet (eng. *effectiveness*) för att se om interventionerna kan användas i ordinarie verksamhet med tillräckligt bibehållna effekter (jfr Flay, Biglan, Boruch, Castro, Gottfredson, Kellam m.fl., 2005; Fraser, Richman, Galinsky & Day, 2009; Fraser & Galinsky, 2010). Fraser med flera (2009) föreslår att denna typ av forskning, utvecklingen av nya interventioner (både deras innehåll och effekter), ska benämnas *intervention research*.

Intervention

I boken används termen intervention som ett samlingsbegrepp för åtgärder, behandlingar, metoder och insatser. Några kännetecken på en intervention är att den (1) är en medveten åtgärd för att åstadkomma en förändring, (2) syftar till att uppnå ett visst mål för en person, familj, skola eller ett samhälle (t.ex. minska eller förebygga psykiska eller sociala problem), (3) sammanfattas i form av en skriftlig eller muntlig överförbar kunskap samt (4) görs tillgänglig genom utbildning, undervisning, handledning eller självstudier (jfr Centers for Disease Control and Prevention, 2007; Fraser m.fl., 2009).

Av ovanstående följer att syftet med interventionen, det vill säga det som interventionen är tänkt att förändra, behöver vara i fokus för utvärderingen. Det kan visserligen vara intressant att veta att en missbruksbehandling har effekter på självkänsla, men mer centralt är om interventionen minskar missbruk. Vidare måste innehållet i interventionen kunna beskrivas för andra, något som i allmänhet förutsätter någon form av skriftlig dokumentation av interventionen. Om man inte kan beskriva det som utvärderas blir utvärderingen meningslös eftersom andra professionella inte har möjlighet att lära sig interventionen.

Evidens

I vardagligt språk likställs evidens ofta med bevis eller fakta som ligger till grund för att något är sant (www.dictionary.com). I ve-

tenskapliga sammanhang betraktas evidens snarare som något explicit, systematiskt (kodifierat med transparenta metoder) och replikerbart (Lomas, Cuyler, McCutcheon, McAuley & Law, 2005).

Alla empiriska observationer kan sägas ge evidens; observationer från en professionell ger en typ av evidens, medan en vetenskaplig utvärdering ger en annan. Båda typer av observationer kan ge viktig information, beroende på frågeställning. Samtidigt har båda begränsningar. Tillförlitligheten i personliga observationer påverkas av människans begränsade kognitiva förmåga. Exempelvis påverkar våra förväntningar oss att se det vi förväntar oss att se. Om man tror att en viss behandling är den bästa kommer man troligen att i högre grad notera tecken som bekräftar det och bortse från tecken som inte gör det. Den som får ta del av en intervention kan också påverkas: tron att en behandling är effektiv kan ha en läkande verkan (placeboeffekt), liksom motsatsen (noceboeffekt). Ett annat problem är att den uppmärksamhet som följer av att delta i en vetenskaplig undersökning kan öka deltagarens prestationer (Hawthorne-effekt). En positiv förändring bland dem som genomgått en behandling kan alltså bero på att de medverkat i en vetenskaplig studie och inte på att behandlingen är effektiv.

Även välgjorda vetenskapliga utvärderingar har begränsningar. Eftersom resultat från effektutvärderingar endast beskriver interventioners genomsnittliga effekter är det inte självklart att interventionen fungerar lika för alla. Det finns ofta modererande variabler som medför att effekten varierar.

Graden av evidens för en viss interventions effekter kan variera med utfallsmått, grupp, kontext och problematik. Det är därför viktigt att beskriva för vem en viss intervention har effekt och under vilka omständigheter.

Utvärderingens design har betydelse för resultatet. Det finns flera exempel på att bättre kontrollerade utvärderingar genererar lägre effekter än utvärderingar med sämre kontroll. Exempelvis undersökte Weisburd och kollegor (2001) 308 utvärderingar av interventioner mot kriminalitet. Utvärderingarna kategoriserades med en

femgradig skala för att gradera vetenskaplig säkerhet, med randomiserat kontrollerade experiment som det ”bästa” alternativet och sambandsstudier som det ”svagaste”. Resultaten visar att ju lägre säkerhet på utvärderingsdesignen, desto effektivare framstår interventionen. En förklaring är att ”svagare” designer inte kontrollerade för sådana förändringar som sker oberoende av interventionen (t.ex. mognad eller byte av umgänge) utan att dessa förändringar enbart tillskrevs interventionen.

Det spelar också roll vad interventionen har jämförts med. Om en intervention jämförs med ingen intervention (eller väntelista) blir effekterna i allmänhet större än om jämförelsen görs med en annan aktiv intervention (Grissom, 1996; Magill & Ray, 2009; Shadish, 2011). Knappast någon intervention är effektiv för alla som tillhör målgruppen. Detta innebär att om två interventioner jämförs och resultaten visar att intervention X har bättre effekt än intervention Y, så betyder det inte att intervention X är tillräckligt bra. För det krävs andra analyser, till exempel att interventionen inte har skadliga effekter för någon grupp.

Sättet att hantera bortfall av deltagare i en utvärdering spelar roll för vilka effekter som en intervention uppvisar. Den strategi som förordas i dag (Wright & Sim, 2003) – att alla som påbörjar en effektutvärdering ska ingå i analyserna (eng. *Intention to treat analysis*, *ITT*) – resulterar ofta i svagare effekter än den strategi som tidigare var vanlig – att endast medta dem som fullföljt behandlingen (eng. *Treatment on the treated*, eller *Treatment of treated, TOT*).

Nya studier tillkommer ständigt, liksom nya interventioner, vilket kan förändra slutsatser om en viss interventions evidens. Inom medicinen uppskattas hälften av dagens kunskapsmassa vara inaktuell, felaktig eller irrelevant om tio år (Nordenström, 2009). Det betyder att en interventions vetenskapliga stöd förmodligen förändras över tid. Andra överväganden som inte handlar om evidens måste också beaktas i kliniska beslut. En intervention ska också vara lämplig i etiska, praktiska (lokala) och kostnadsmässiga perspektiv.

Kausalitet[3]

Effektutvärderingar handlar om att avgöra om en specifik intervention påverkar utfallet för en viss grupp individer. Det handlar således om att bedöma orsakssamband, det vill säga kausalitet. Det finns en rad krav för tolkningen av kausalitet. Dessa krav är nödvändiga, men garanterar inte att kausalitet faktiskt föreligger.

Ett första krav är att interventionen föregår effekten. Om orsak och effekt undersöks samtidigt går det inte att avgöra om det finns ett orsakssamband. Om effekten föregår den tilltänkta orsaken vet man däremot säkert att denna inte kan ha orsakat effekten.

Ett andra krav är att det måste finnas ett samband mellan orsak och effekt; när orsaken förekommer ska också effekten finnas, och när orsaken inte finns ska inte heller effekten uppträda. Inom samhällsvetenskap förekommer dock sällan absoluta samband eftersom människor påverkas av många parallella faktorer. När orsaken förekommer visar sig effekten ibland, ibland inte. Ibland leder missbruksbehandling till att missbrukare rehabiliteras, ibland inte. På samma sätt kan effekter uppträda utan att den tänkta orsaken förekommit – missbrukare upphör exempelvis ibland med sitt missbruk på egen hand och utan missbruksbehandling. Även med starka samband finns ingen garanti att den studerade interventionen faktiskt orsakat effekten. Det kan finnas andra faktorer som forskaren inte kontrollerat för som orsakar att effekten uppträder. För att kunna avgöra om det finns ett samband som är tillräckligt starkt för att betraktas som sant (och inte orsakat av slumpen) används statistik.

Ett tredje krav är att en experimentell manipulation kan förutse en förändring. Dessa tre krav är inbyggda i den säkraste typen av effektutvärderingar.

Ett fjärde krav är att olika studier av olika forskare från olika platser ska generera konsistenta resultat. På samma sätt stärks slutsatserna om olika teoretiskt relaterade utfallsmått i en och samma ut-

3 Se även www.drabruzzi.com/hills_criteria_of_causation.htm

värdering pekar i samma riktning, exempelvis om en intervention mot ungdomars antisociala beteenden visar på ett minskat drogbruk, minskad kriminalitet och minskat skolk.

Ett femte krav är att sambandet mellan en intervention och ett utfall är teoretiskt rimligt och att det inte finns andra rimliga förklaringar till sambandet. Detta är viktigt, eftersom det snarast är regel än undantag inom samhällsvetenskaplig forskning att flera möjliga förklaringsfaktorer samvarierar.

Taxonomi för effektutvärderingar

Det finns ingen enhetlig terminologi för effektutvärderingar.[4] Man brukar dock skilja mellan experiment och observationsstudier (figur 1:2). Experimentet är en studie där forskaren medvetet introducerar en intervention för att avläsa dess effekter (Shadish, Cook & Campbell, 2002). I allmänhet innebär det att förändringen avläses genom en mätning före respektive efter interventionen. Interventionen kan introduceras av forskaren själv eller av exempelvis professionella i reguljär verksamhet. Syftet är att studera en viss variabels påverkan på en annan variabel i kausala termer.

I en observationsstudie (icke-experimentell studie; icke-interventionsstudie; naturalistisk studie) ingriper inte forskaren i ett händelseförlopp utan följer och observerar naturligt förekommande händelser i människors liv. Syftet är fortfarande att försöka tolka kausala samband men där det är etiskt eller praktiskt omöjligt att intervenera (Cochran, 1965).

4 www.ahrq.gov/clinic/out2res/outcom1.htm; Jefford m.fl. (2003).

Experimentella studier

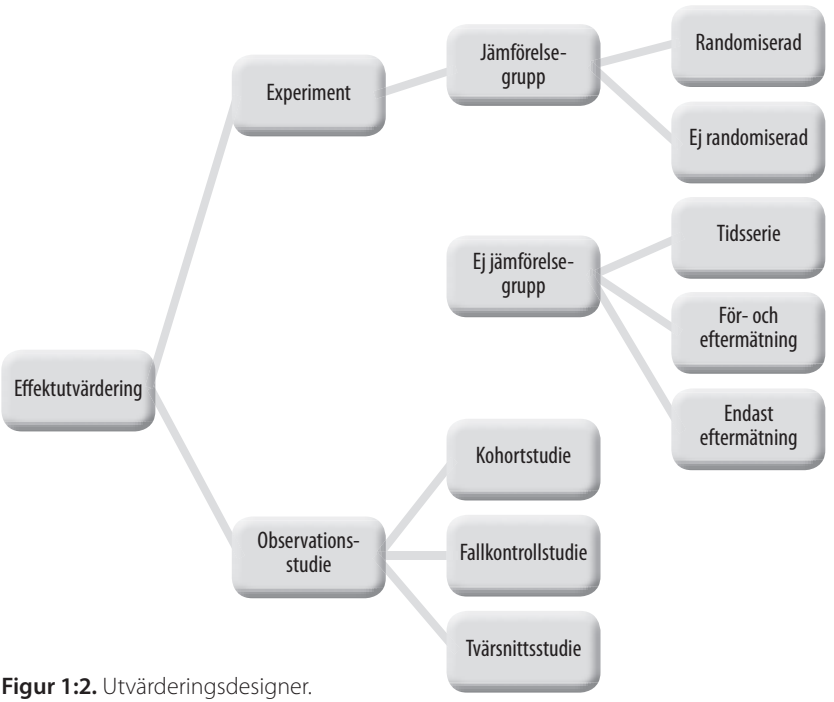
Om experimentet är en kontrollerad studie finns en jämförelsegrupp (även kallad kontrollgrupp)[5] som får en annan intervention än experimentgruppen (t.ex. standardbehandling), placebobehandling eller ingen behandling alls och vars resultat jämförs med experimentgruppens. Denna jämförelsegrupp underlättar slutsatser om orsakssamband eftersom den skapar en kontroll för andra möjliga förklaringar till resultaten, till exempel att förbättringar beror på en placebo- eller en Hawthorne-effekt. Utan jämförelsegrupp är det svårt att avgöra om en förändring för en grupp beror på interventionen eller på andra saker (t.ex. att det normala är att människor förändras). Jämförelsegruppen möjliggör en kontrafaktisk analys där det som är kontrafaktiskt är motsatsen till fakta. I det här fallet handlar det om vad som skulle ha hänt samma individer om de inte hade fått interventionen. Per definition är det omöjligt att studera eftersom interventionen ju inträffat. Man försöker därför skapa estimat på det kontrafaktiska, vanligtvis genom ett alternativ som inte involverar interventionen.

I en randomiserad kontrollerad studie (eng. *randomised controlled trials, RCT*) fördelas deltagarna slumpvis på experiment- och jämförelsegrupperna. Det skapar i teorin grupper som i allt väsentligt är lika varandra i genomsnitt. Det enda som kommer att skilja grupperna åt systematiskt är vilken intervention de får. Om studien är experimentell med jämförelsegrupp men där deltagarna inte randomiseras till experiment- och jämförelsegrupperna handlar det om en icke-randomiserad experimentell studie. Denna designtyp kallas även kvasiexperimentell (Shadish m.fl., 2002).

Det finns tre typer av experiment som saknar jämförelsegrupp:

---

5 Exempelvis http://medical-dictionary.thefreedictionary.com/controlled+trial *Kontrollgrupp* och *jämförelsegrupp* används i dag i allmänhet som synonymer. Tidigare användes kontrollgrupp för att beteckna ett kontrafaktiskt alternativ som inte fått någon behandling och jämförelsegrupp som de som fått en annan behandling, till exempel standardbehandling (Shadish m.fl., 2002).



Figur 1:2. Utvärderingsdesigner.

tidsserier (eng. *time serie*)[6], för- och eftermätning (eng. *one-group pre-test-posttest*) samt endast eftermätning (eng. *one-group posttest-only*). Endast eftermätning utan jämförelsegrupp är en mycket svag utvärderingsdesign. Utan förmätning är det svårt att veta om det skett någon förändring, och frånvaron av jämförelsegrupp gör det svårt att veta vad som skulle ha skett utan behandling. Det finns dock exempel där designen använts framgångsrikt. Ett exempel som nämns av Shadish med flera (2002) är från den kanadensiska provinsen Ontario. År 1966 beslöt provinsen att försöka identifiera och behandla alla barn som föddes med PKU (brist på ett enzym som kan leda till en allvarlig hjärnskada). En uppföljning visade att 44 av 45 barn som behandlats inte visade tecken på utvecklingsstörning. Statistik

6 En tidsseriestudie kan också vara ”kontrollerad” om det finns en jämförelsegrupp.

från åren före 1966 visade betydligt högre andelar barn som utvecklat utvecklingsstörning.

Genom att inkludera en mätning före behandlingen får man viss information om vad som skulle kunna ha hänt undersökningsgruppen om behandlingen inte skett. Trots det ger denna design med för- och eftermätning dålig kontroll över om en uppmätt förändring beror på interventionen eller på andra faktorer. Tidsserier, det vill säga upprepade mätningar före och efter interventionen, ger möjlighet att utläsa trender och cykliska mönster. Om en rad förmätningspunkter ger en stabil problemnivå som förändras radikalt efter interventionen (t.ex. genom att successivt förbättras) kan det ge argument för kausala effekter.

Randomiserade kontrollerade utvärderingar är den studietyp som ger teoretiskt bästa möjligheten att uttala sig om en intervention orsakat effekter eftersom:

- experimentet gör att effektvariablen kan mätas före intervention och möjliga effekter
- kontrollgruppen ger möjlighet till kontrafaktisk inferens
- randomiseringen minskar risken för selektionseffekter.

Även icke-randomiserade kontrollerade studier kan ge möjlighet att dra kausala slutsatser, framför allt efter statistisk kontroll för initiala skillnader (Shadish m.fl., 2002). Shadish, Clark och Steiner (2008) åskådliggör detta i en studie där de randomiserade 445 studenter till två betingelser. I den första kom studenterna sedan att randomiseras till träning av matematik eller träning av vokabulär. Den andra betingelsen innebar att studenterna själva fick välja mellan matematik eller vokabulär. Den första betingelsen är med andra ord RCT och den andra ett icke-randomiserat experiment där studenterna självselekterat till de två behandlingarna. I RCT-betingelsen kom studenterna att fördelas ungefär jämnt mellan vokabulär och matematik, medan det var väsentligt fler som valde vokabulär när studenterna fick välja själva. Dessutom lät forskarna studenterna besvara en mängd frågor innan träningen påbörjades.

Eftersom samma träningsmetoder användes för båda betingelserna undersöker studien designens betydelse för resultaten. RCT-betingelsen kan sägas ge den sanna effekten av träningen medan icke-RCT-betingelsen ger ett som snedvridits på grund av selektion. När resultaten i den senare betingelsen justerades statistiskt med hjälp av 25 relevanta kovariater (t.ex. ”mattefobi”) kunde ”bias” minskas med cirka 90 procent. Det krävs dock minst 500 individer för att nå detta resultat; med färre ökar felet exponentiellt (Shadish, 2011). Dessutom krävs att de relevanta kovariaterna är mätta, vilket är en utmaning med tanke på den relativt begränsade kunskap vi har om vad som är viktigt. Med rätt förutsättningar går det med andra ord att få trovärdiga resultat även med icke-randomiserade kontrollerade studier.

Vid experimentella utvärderingar skiljer man vidare mellan modellutvärderingar (eng. *efficacy trials*) och verksamhetsbaserade utvärderingar (eng. *effectiveness trials*). Modellutvärderingar är experimentella utvärderingar där forskarna har maximal kontroll över att interventionen implementerats korrekt. I allmänhet är det forskaren själv eller dennes personal som ansvarar för behandlingen. I verksamhetsutvärderingen studeras hur interventionen fungerar i vardaglig verksamhet där ordinarie personal ansvarar för behandlingen. Skillnaden mellan de två typerna av utvärderingar är inte bara teoretisk, utan har också resulterat i signifikant högre effekter i modellutvärderingar (Curtis, Ronan & Borduin, 2004; Lösel & Beelmann, 2003; Petrosino & Soydan, 2005).

Observationsstudier

En observationsstudie (eng. *observational study*) syftar också till att dra kausala slutsatser, men där forskaren inte har implementerat eller randomiserat en intervention till deltagare.[7] Termen observation

7 Ett sätt att förstå skillnaden mellan experiment och observationsstudier är inte förekomsten av manipulation utan snarare om manipulationen förekommer på ett kontrollerat eller okontrollerat (osystematiskt) sätt (personlig kommunikation, William Shadish, 2012-01-03).

syftar alltså inte på metoden för datainsamling (som kan vara exempelvis ett test eller en enkät), utan på forskarens mer passiva roll.

Observationsstudier omfattar tre typer av design: kohortstudie (eng. *cohort study*), fall-kontrollstudie (eng. *case-control study*) och tvärsnittsstudie (eng. *cross-sectional study*). I en kohortstudie följer man en grupp individer över tid och samlar data om vissa förmodade orsaksfaktorer (t.ex. rökning) och utfall (t.ex. lungcancer) vid flera mättillfällen. I en fall-kontrollstudie med samma syfte studerar man andelen rökare bland individer med lungcancer (fall) och jämför den med andelen rökare bland personer utan lungcancer (kontroll). I en tvärsnittsstudie samlar man in data från en grupp individer vid en enda tidpunkt och studerar samband, exempelvis mellan förekomsten av rökning och lungcancer. Observationsstudier kan inkludera jämförelsegrupper.

Möjligheterna att dra pålitliga slutsatser om interventionens effekter är mindre med ett okontrollerat experiment och med en observationsstudie. Skillnaden mellan en kontrollerad experimentell studie och observationsstudie som omfattar flera granskade grupper är att forskaren i den kontrollerade experimentella studien medvetet väljer en viss interventions- och jämförelsegrupp enligt på förhand fastställda principer medan forskaren i en observationsstudie studerar grupper som uppstår på ett naturligt sätt (t.ex. kvinnor i en region som genomgått mammografi). Det centrala problemet med en observationsstudie är att man inte vet principen för att personer fått en viss intervention men att man bör utgå från att den inte är slumpmässig (West & Thoemmes, 2010).

Observationsstudier omfattar ofta stora datamaterial och används inom epidemiologisk forskning för att studera samband mellan framför allt sjukdomar och deras tänkbara orsaker. De är också viktiga informationskällor för bland annat identifikation av långtidseffekter och sällsynta biverkningar (Black, 1996). I de fall en faktor har en så tydlig inverkan på ett utfall att alternativa förklaringar med stor säkerhet kan uteslutas, kan observationsstudier även tillhandahålla pålitlig information om orsak och verkan. Det vore till exempel både onödigt

och oetiskt att genomföra randomiserade kontrollerade studier för att bekräfta att tobaksrökning är skadligt för hälsan.

Populationsbaserade folkhälsointerventioner kan vara speciellt svåra att utvärdera i experimentella studier. Till svårigheterna hör att uppnå rimlig statistisk styrka när analysenheten är en viss kommun eller region, problem att kontrollera att interventionen inte sprids via elektronisk kommunikation (kontaminering), behov av långa uppföljningstider för att ha rimlig möjlighet att upptäcka effekter, svårigheter att inhämta informerat samtycke och stora kostnader som är förknippade med folkhälsointerventioner (Sanson-Fischer, Bonevski, Green & D'Este, 2007).

Steg i en effektutvärdering

Effektutvärderingar inkluderar ett antal överväganden och beslut. Därför behövs en genomtänkt forskningsplan som säkrar att alla viktiga delar beaktats och som vägleder genomförandet. Man kan tala om nio steg (tabell 1:1), varav denna bok fokuserar på några av stegen (jfr HM Treasury UK, 2011).

Det första steget är att definiera målet med utvärderingen: Vad ska interventionen åstadkomma och för vilka? Hur är interventionen kopplad till resultat och effekter (förändringsteorin)? Är det en modellutvärdering eller en utvärdering i ordinarie verksamhet? Vad ska resultaten användas till? Finns det en beställare eller avnämare? Ska resultaten användas för ett speciellt ändamål? Vilken befintlig forskning behöver analyseras för att få vägledning för design, mätinstrument och frågeställningar eller hypoteser? Dessa klargöranden är viktiga för den fortsatta planeringen.

Det andra steget är att identifiera utvärderingsresultatens målgrupp: För vem är resultaten viktiga, till exempel personer inom departement, myndigheter, utbildningar eller kommunala verksamheter (eng. *stakeholder, end-user, policy maker, purveyor*)? Hur ska dessa målgrupper involveras i planering och implementering? Målgruppen kan ge information om resultat som är speciellt viktiga (t.ex.

att kostnadsanalyser behövs), hur allmängiltiga resultaten behöver vara (t.ex. att studien behöver inkludera olika geografiska områden och minoritetsgrupper) och vilket stöd som behövs för implementering (t.ex. beskriva steg i implementeringsarbetet). Om det handlar om en utvärdering i ordinarie verksamhet är det nödvändigt att de som arbetar i den involveras i förankrings- och planeringsarbetet.

Det tredje steget handlar om att definiera frågeställningar eller hypoteser. Vad är prioriterat och vad är sekundärt? Vilka nya frågor besvarar utvärderingen? Vilken kunskap behöver beslutsfattare ha för att kunna ta beslut om att använda metoden?

Det fjärde steget handlar om att välja design. Utöver att utvärdera effekterna av en intervention – behövs exempelvis också en processutvärdering eller en ekonomisk utvärdering? Hur omfattande behöver utvärderingen vara för att ge tillräckligt säkra resultat?

Det femte steget handlar om att identifiera vilka data som ska samlas in. Frågeställningar eller hypoteser styr vilka data som behöver registreras och när. Finns exempelvis en frågeställning om ekonomiska analyser behöver data förmodligen samlas in löpande. Finns det registerdata som kan användas? Val av tidpunkt för när utvärderingen ska genomföras påverkar resultaten. En studie som inleds ”för tidigt” kan medföra att den intervention som utvärderas inte existerar eftersom den inte implementerats tillräckligt väl för att erbjuda ett acceptabelt test. Å andra sidan kan det vara svårt att motivera en utvärdering av en redan etablerad intervention där de professionella är övertygade om att interventionen är effektiv. Val av tidsperiod för när data ska samlas in för att avläsa effekter är också en central fråga. Att välja en kort uppföljningstid är en fördel eftersom det snabbt går att uttala sig om interventionens värde, men det innebär också en risk att effekterna snabbt klingar av. Vissa effekter kan ta år att visa sig och det gör utvärderingen betydligt kostsammare.

Det sjätte steget handlar om behov av ekonomiska och personella resurser. Med för begränsad budget kan det vara oetiskt att starta en utvärdering eftersom de resultat som kan erhållas inte kommer att vara tillförlitliga. Om resurserna är små kan det bli aktuellt att

begränsa studien till undergrupper av målgruppen, ersätta intervjuer med telefonintervjuer eller postenkäter eller begränsa uppföljningstiden. Det är också viktigt att säkra vetenskaplig (t.ex. för avancerade statistiska analyser) och praktisk kompetens (t.ex. specialutbildade datainsamlare). Utöver detta krävs kvalitetskontroll och kvalitetssäkring. Detta innebär att eventuella brister i design, metod, datainsamling med mera identifieras och åtgärdas innan utvärderingen påbörjas och att det säkerställs att effektutvärderingen motsvarar etiska och professionella normer. Det kan göras genom normala forskningsprocesser (t.ex. peer review-förfarande vid ordinarie forskningsmedelsansökningar) eller med hjälp av interna och externa granskare. Fyra faktorer är särskilt viktiga för kvalitetssäkring: (a) forskares oberoende och objektivitet; (b) att centrala målgrupper och intressenter involveras genom styrgrupp eller liknande; (c) transparens (öppenhet) i planering av utvärderingens mål och genomförande samt (d) att effektutvärderingen genomförs med en vetenskaplig standard som ökar sannolikheten att utvärderingen ger tillförlitliga resultat. Det senare är det som denna bok behandlar.

Det sjunde steget handlar om få ett forskningsetiskt godkännande för utvärderingen.

Det åttonde steget är att genomföra datainsamling och analys samt sammanfatta resultat i en vetenskaplig rapport. Viktiga frågor är vem som ansvarar för olika delar, under vilken period datainsamling sker, om datainsamlingsmetoder och datainsamlingen behöver testas samt vem som håller kontakt med styr- och eventuella referensgrupper.

Det sista steget handlar om att sprida resultaten. I vilka medier ska resultaten spridas? Hur ska resultaten spridas och i förekommande fall implementeras? Berörs någon utbildning?

Steg i en effektutvärdering

1. Definiera mål och avsedda resultat

- Vad är interventionens logik (förändringsteori) om hur den orsakar effekter?

2. Definiera målgruppen för utvärderingens resultat

- Vilka är de primära målgrupperna för resultaten och hur blir de involverade?

3. Identifiera relevanta frågeställningar

- Vad behöver beslutsfattare veta om interventionens effekter och hur den implementerats?
- Hur omfattande är utvärderingen?

4. Val av utvärderingsdesign

- Ska utvärderingen belysa effekter, processer eller en kombination?
- Behövs en ekonomisk utvärdering?
- Hur omfattande behöver utvärderingen vara?

5. Identifiera krav på data

- Vilka data har redan samlats in och vad behöver ytterligare samlas in?
- Vid vilka tidpunkter bör effekterna mätas?
- Vem ansvarar för datainsamling och vilken organisation behövs?
- Hur ska data registreras och lagras säkert?

6. Identifiera nödvändiga resurser och styrmekanismer

- Vilken budget behövs?
- Vem är projektledare, vem ansvarar för datainsamling, analys och rapportskrivning?
- Vilka ingår i styrgruppen?
- Hur sker kvalitetssäkring?

7. Säkra forskningsetiskt godkännande

8. Genomföra utvärderingen

- Behöver datainsamlingsinstrument testas eller datainsamlingen pilotas?
- När startar och slutar utvärderingen?

9. Spridning av utvärderingsresultaten

- Hur ska resultaten spridas och i förekommande fall implementeras?
- Hur ska resultaten matchas till kunskapsstyrningshjulet (figur 1:3)?

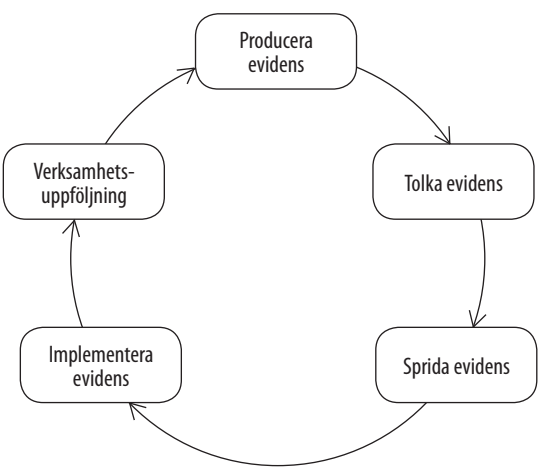
Kunskapsstyrning

Många länder, inklusive de nordiska, prioriterar allt mer ett systematiskt utvärderingsarbete av psykosociala och pedagogiska interventioner. I Norge är det bland annat tydligt vad gäller barn och unga med allvarliga beteendeproblem, där Atferdssenteret spelar en viktig roll (Ogden, Amlund-Hagen, Christensen & Askeland, 2009). I Sverige har olika regeringar sedan länge haft ambitionen att göra exempelvis socialt arbete forskningsbaserat. Inrättandet av Centrum för utveckling av socialt arbete (CUS) år 1992 och Institutet för utveckling av metoder i socialt arbete (IMS) år 2004 är konkreta exempel på detta. Exempel från psykiatrin är Nationellt centrum för suicidforskning och prevention av psykisk ohälsa (NASP) samt Centrum för Evidensbaserade Psykosociala Insatser (CEPI).

Kunskapsstyrning handlar om att hjälpa människor att fatta välinformerade beslut inom politik och förvaltning, baserat på bästa tillgängliga vetenskapliga kunskap. Den engelska term som brukar användas är *evidence-based policy*. Det handlar exempelvis om att identifiera vilka interventioner som har bästa vetenskapliga evidens om önskade effekter så att beslutsfattare har bästa möjliga underlag för beslut. Däremot innebär det inte att forskare eller forskningsresultat ska ersätta politiker och andra beslutsfattare. Forskning utgör en av flera informationskällor och konkurrerar med ideologi, politisk övertygelse, expertutlåtanden, ekonomi, professionella normer och kultur (Nutley, Walter & Davies, 2007). Det kan handla om att förändra utbudet av interventioner, utmönstra det som visat sig vara skadligt eller som saknar effekt, att införa en ny och mer effektiv åtgärd eller att förhindra att en ny metod införs i verksamheten när den saknar evidens om positiva effekter.

Kunskapsstyrning består av alla styr- och ledningsprocesser som bidrar till att etablera och implementera bästa tillgängliga kunskap och som bygger en ändamålsenlig infrastruktur för kunskapsbildning och kunskapsutveckling (Socialstyrelsen, 2011b). Kunskapsstyrning kan delas in i fem sammanlänkade led. Det första ledet är

att producera vetenskaplig kunskap (evidens), företrädesvis genom effektutvärderingar (figur 1:3). Det görs av forskare inom universitet och högskolor, inom forsknings- och utvecklingsenheter, privata företag och till en del också inom myndigheter som Socialstyrelsen. Det andra ledet handlar om att tolka den tillgängliga evidensen, exempelvis genom systematiska översikter. Det kan göras av forskare på universitet, myndigheter och av massmedia. Det tredje ledet är att sprida information om evidens till beslutsfattare och professionella. Samma aktörer som tidigare räknats upp kan ansvara för detta. Det fjärde ledet är att implementera evidens i praktiskt arbete. Enligt ny implementeringsforskning (t.ex. Fixsen, Naoom, Blasé, Friedman & Wallace, 2005) bör det ske i den lokala verksamheten och där den verksamhetsansvarige chefen har ett stort ansvar. Det femte och sista ledet handlar om att följa upp verksamheten för att skapa klarhet i om den fungerar eller behöver modifieras. Denna uppföljning och problemformulering kan ske lokalt, regionalt eller nationellt. Denna kunskap går sedan tillbaka till forskare som utvecklar nya interventioner etcetera.



Figur 1:3. Kunskapsstyrning (jfr Shepherd, 2007).

Den forskare som vill att hans eller hennes forskning ska komma till praktisk användning behöver anpassa sig till de krav som utvecklats för att producera och tolka evidens. Om så inte sker är risken stor att kunskapen aldrig lämnar akademin. En genomgång av fyra systematiska översikter från Cochrane och Campbell Collaboration (Binks, Fenton, McCarthy, Lee, Adams & Duggan, 2006; Cameron, Murray, Gillespie, Robertson, Hill, Cumming m.fl., 2010; Furlong, McGilloway, Bywater, Hutchings, Smith & Donnelly, 2012; Zwi, Jones, Thorgaard, York & Dennis, 2012) visar exempelvis att endast 15 procent av sammanlagt 454 relevanta utvärderingar medtogs i analyserna. Vanliga orsaker till att utvärderingar exkluderades var att de saknade eller hade fel kontrollbetingelse, att undersökningsgruppen inte motsvarade de uppställda kriterierna, att interventionen inte motsvarade kriterierna samt att relevanta effektmått saknades.

Syntes av effektutvärderingar med hjälp av systematiska översikter

En central del i kunskapsstyrning är tillgång på vetenskaplig kunskap om interventioners effekter. Den systematiska översikten (eng. *systematic review*) är en strukturerad metod för att identifiera, välja ut och sammanfatta forskning om en definierad fråga om interventioners effekter eller diagnostiska metoders effekt och kvalitet (jfr Anttila, 2007; Eliasson, 2000). Forskningsöversikter i sig är inget nytt utan utgör en traditionell del i forskning; de flesta vetenskapliga artiklar omfattar en översikt av kunskapsläget.

Begreppet systematisk översikt härrör från en artikel av Mulrow (1987). Hon granskade 50 traditionella översiktsartiklar som publicerats i fyra välmeriterade medicinska tidskrifter. Ingen av översikterna uppfyllde alla uppställda kriterier på en god vetenskaplig syntes av kunskap. Den systematiska översikten är inriktad på en fokuserad fråga och följer samma vetenskapliga metodik som primärforskning. Den ska redovisa frågeställning, metod, genomförande och tolkning på ett sådant sätt att andra forskare kan reproducera resultaten. Jämförelser mellan traditionella översikter (eng.

narrative reviews[8]) och systematiska översikter inom medicin visar att traditionella översikter ökar risken för systematiska fel (eng. *bias*) med över- eller underskattning av behandlingseffekter (t.ex. Antman, Lau, Kupelnick, Mosteller & Chalmers, 1992).

Enligt Anttila (2007) finns några kännetecken på den systematiska översikten:

1. Målgruppen består i första hand av beslutsfattare, praktiker och brukare och först i andra hand av forskare. Detta beror på att målet med systematiska översikter är att bidra till ett vetenskapligt beslutsunderlag i en evidensbaserad praktik.
2. Syftet är att väga samman forskningsresultat från alla kända och relevanta primärstudier som håller acceptabel vetenskaplig kvalitet, inklusive så kallad grå litteratur (utvärderingar som inte är publicerade i vetenskapliga tidskrifter). Då det är lämpligt vägs resultaten samman statistiskt i en metaanalys.
3. Den systematiska översikten är ett levande dokument som ska revideras regelbundet. Revideringar bör göras när nya forskningsresultat tillkommer, men även när brister i översikten uppdagas. Tidpunkten för nästa planerade revidering finns normalt sett angiven i dokumentet.
4. Samtliga arbetsmoment genomförs enligt ett protokoll och det finns explicita kriterier för inklusion och exklusion av studier. Viktiga beslut och åtgärder under arbetsprocessen dokumenteras för att det ska vara möjligt för den kritiske läsaren att värdera risken att slutsatserna snedvridits. Med hjälp av protokollet ska det vara möjligt att replikera resultaten för en oberoende granskare.

Viktiga faktorer för att bedöma evidensstyrka är studiekvalitet, samstämmighet mellan olika studier, överförbarhet från en kontext till

8 Narrativa översikter ger ofta en övergripande beskrivning av ett ämne snarare än att söka besvara en specifik fråga, som hur effektiv en viss intervention är för ett visst tillstånd (problem). Narrativa översikter redovisar sällan hur sökningen efter litteratur gick till eller vilka kriterier som använts för att välja ut det som inkluderats.

ett annan, precision i data och risk för publikationsbias.[9] Informationen sammanvägs för att gradera trovärdigheten i det vetenskapliga underlaget. Det system som allt flera använder är det internationellt utvecklade The Grading of Recommendations Assessment, Development and Evaluation (GRADE).[10] GRADE:s evidensgradering bygger på en fyrgradig skala från starkt, måttligt, svagt till mycket svagt vetenskapligt underlag.

De viktigaste organisationerna för systematiska översikter internationellt är Cochrane Collaboration och Campbell Collaboration. I Sverige görs systematiska översikter av Statens beredning för medicinsk utvärdering (SBU) och enheten för kunskapsöversikter inom Socialstyrelsen. GRADE används i Sverige av SBU och Socialstyrelsen och internationellt av Cochrane Collaboration, National Institute of Health and Clinical Excellence (NICE) och WHO.

Fördjupningslitteratur

Anttila, S. (2008). Systematiska översikter. Ingår i U. Jergeby (red.), *Evidensbaserad praktik i socialt arbete* (sid. 88–110). Stockholm: Gothia Förlag.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., m. fl.. (2005). Standards of evidence. Criteria for efficacy, effectiveness and dissemination. *Preventions science, 6,* 151–175.

Fraser, M. W., Richman, J. M., Galinsky, M. J. & Day, S. H. (2009). *Intervention research. Developing social programs.* New York: Oxford university press.

Lilienfeld, S. O. (2007). Psychological treatments that cause harm. *Perspectives on psychological science, 2*, 53–70.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.

Shepherd, J. (2006). The production and management of evidence for public service reform. *Evidence & policy, 3,* 231–251.

9 http://www.sbu.se/sv/Evidensbaserad-vard/Utvardering-av-metoder-i-halso-och-sjukvarden--En-handbok/

10 http://www.gradeworkinggroup.org/

Referenser

Antman, E. M., Lau, J., Kupelnick, B., Mosteller, F. & & Chalmers, T. C. (1992). A comparison of results of meta-analyses of randomized controlled trials and recommendations of clinical experts. Treatment of myocardial infarctions. *JAMA, 268,* 240–248.

Anttila, S. (2008). Systematiska översikter. Ingår i U. Jergeby (red.), *Evidensbaserad praktik i socialt arbete* (s. 88–110). Stockholm, Gothia.

Beutler, L. E. (2000). David and Goliath. When empirical and clinical standards of practice meet. *American Psychologist, 55*, 997–1007.

Binks, C., Fenton, M., McCarthy, L., Lee, T., Adams, C. E. & Duggan, C. (2006). Psychological therapies for peoplewith borderline personality disorder. *Cochrane Database of Systematic Reviews, 1.* Art. No.: CD005652. DOI: 10.1002/14651858.CD005652.

Black, N. (1996). Why we need observational studies to evaluate the effectiveness of health care. *British Medical Journal, 312*, 1215–1218.

Cameron, I. D., Murray, G. R., Gillespie, L. D., Robertson, M. C., Hill, K. D., Cumming, R. G. m.fl. (2010). Interventions for preventing falls in older people in nursing care facilities and hospitals. *Cochrane Database of Systematic Reviews 1*. Art.No.: CD005465. DOI: 10.1002/14651858.CD005465.pub2.

Centers for Disease Control and Prevention (2007). Improving public health practice through translational research. Nedladdat 2010-04-04 från http://grants.nih.gov/grants/guide/rfa-files/rfa-cd-07-005.html

Chaffin, M. & Friedrich, B. (2004). Evidence-based treatments in child abuse and neglect. *Children and Youth Services Review, 26*, 1097–1113.

Cochran, W. G. (1965). The planning of observational studies of human populations (with Discussion). *Journal of the Royal Statistical Society. Series A, 128*, 134–155.

Curtis, N. M., Ronan, K. R. & Borduin, C. M. (2004). Multisystemic treatment: A meta-analysis of outcome studies. *Journal of Family Psychology, 18,* 411–419.

Eliasson, M. (2000). Den systematiska översikten grundval i evidensbaserad medicin. *Läkartidningen*, 97 (22), 2726–2728.

Fixsen, D. L., Naoom, S. F., Blasé, K. A., Friedman, R. M. & Wallace, F. (2005). *Implementation research: A synthesis of the literature*. Tampa, Florida: University of South Florida, Louise de la Parte Florida Mental Health Institute, The National Implementation Research Network.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S. m.fl. (2005). Standards of evidence. Criteria for efficacy, effectiveness and dissemination. *Preventions science, 6,* sid. 151–175.

Fraser, M. W. & Galinsky, M. J. (2010). Steps in intervention research: designing and developing social programs. *Research on social work practice, 20*, 459–466.

Fraser, M. W., Richman, J. M., Galinsky, M. J. & Day, S. H. (2009). *Intervention research. Developing social programs.* New York: Oxford University Press.

Furlong, M., McGilloway, S., Bywater, T., Hutchings, J., Smith, S.M. & Don-

nelly, M. (2012). Behavioural and cognitive-behavioural group-based parenting programmes for early-onset conduct problems in children aged 3 to 12 years. *Cochrane Database of Systematic Reviews,* 2. Art. No.: CD008225. DOI: 10.1002/14651858.CD008225.pub2.

Greenwood, P. (2008). Prevention and intervention programs for juvenile offenders: the benefit of evidence-based practive. *The future of children, 18*, 11–36.

Grissom, R. J. (1996). The magical number .7 + .2: Meta-meta-analysis of the probability of superior outcome in comparisons involving therapy, placebo, and control. *Journal of Consulting and Clinical Psychology, 64*, 973–982.

HM Treasury UK (2011). The Magenta Book - Guidance for evaluation. Available online PDF at: http://bit.ly/v9642L

Jefford, M., Stockler, M. R. & Tattersall, M. H. N. (2003). Outcomes research: what is it and why does it matter? *Internal medicine journal, 31,* 110–118.

Kennedy, S. S., Mercer, J., Mohr, W. & Huffine,C. W. (2002). Snake oil, ethics, and the First Amendment: What´s a profession to do? *American Journal of Orthopsychiatry, 72,* 5–15.

Lilienfeld, S. O. (2007). Psychological treatments that cause harm. *Perspectives on psychological science, 2*, 53–70.

Lomas, J., Cuyler, T., McCutcheon, C., McAuley, L. & Law, S. (2005). *Conceptualization and combining evidence for health system, guidance. Canadian health service research foundation*. Ontario: Ottowa.

Lösel, F. & Beelmann, A. (2003). Effects of child skill training in preventing antisocial behaviour: a systematic review of randomized evaluations. *The Annals of the American Academy,* 587: 84–109.

Magill, M. & Ray, L. A. (2009). Cognitive-behavioral treatment with adult alcohol and illicit drug users: a meta-analysis of randomized controlled trials. *Journal of Studies on Alcohol and Drugs, 80,* 516–527.

Mulrow, C. D. (1987). The medical review article: state of the science. *Ann Intern Med, 106,* 485–488.

Nordenström, J. (2009). *Evidensbaserad medicin i Sherlock Holmes fotspår.* Stockholm: Karolinska Institutet University Press.

Nutley, S. M., Walter, I. & Davies, H. T. O. (2007). *Using Evidence: How Research Can Inform Public Services*. Bristol: The Policy Press.

Oakley, A. (1998). Experimentation and social interventions: a forgotten but important history. *Education and debate, 317*, 1239–1242.

Ogden, T., Amlund-Hagen, K., Askeland, E. & Christensen, B. (2009). Implementing and evaluating evidence-based treatments of conduct problems in children and youth in Norway. *Research on Social Work Practice, 19,* 582–591.

Petrosino, A. & Soydan, H. (2005). The impact of program developers as evaluators on criminal recidivism: Results from a meta-analysis of experimental and quasi-experimental research. *Journal of Experimental Criminology,* 1, 435–450.

Sanson-Fischer, R. W., Bonevski, B., Green, L. W. & D'Este, C. (2007). Limitations of the randomized controlled trial in evaluating population-based health

interventions. *Am. J. Prev. Med., 33*, 155–161.
Shadish, W. R. (2011). Randomized Controlled Studies and Alternative Designs in Outcome Studies: Challenges and Opportunities. *Research on social work practice, 21,* 636–643.
Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.
Shadish, W. R., Clark, M. H., Steiner, P. M. (2008). Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments. *Journal of the American Statistical Association. 103,* 1334–44.
Shepherd, J. (2006). The production and management of evidence for public service reform. *Evidence & policy, 3,* 231–251.
Singh, S. & Ernst, E. (2008). *Salvekvick och kvacksalveri. Alternativmedicin under luppen*. Stockholm: Leopard förlag.
Socialstyrelsen (2009a). *Socialtjänstens öppna verksamheter för barn och unga – en nationell inventering av metoder*. Stockholm: Socialstyrelsen.
Socialstyrelsen (2009b). *Skolans metoder för att förebygga psykisk ohälsa hos barn. En nationell inventering i grundskolor och gymnasieskolor*. Stockholm: Socialstyrelsen.
Socialstyrelsen (2009c). *Förskolans metoder för att förebygga psykisk ohälsa hos barn – en nationell inventering*. Stockholm: Socialstyrelsen.
Socialstyrelsen (2011a). *Svensk och internationell forskning om sociala interventioners effekter*. Stockholm: Socialstyrelsen.
Socialstyrelsen (2011b). *På väg mot en evidensbaserad praktik inom socialtjänsten. Kartläggning, analys och förslag för att förbättra kunskapsstyrning*. Stockholm: Socialstyrelsen.
Stanley, J. C. (1957). Controlled experimentation in the classroom. *Journal of experimental education, 25,* 195–201.
Sundell, K. & Stensson, E. (2010). *Förekomst av effektutvärderingar i doktorsavhandlingar*. Stockholm: Socialstyrelsen.
Sundell, K. & Vinnerljung, B. (2008). Goda intentioner kan vålla skada – om etik i socialt arbete. I U. Jergeby (red.), *Evidensbaserad praktik i socialt arbete* (s. 36–46). Stockholm: Gothia Förlag/IMS.
Weisburd, D., Lum, C. M. & Petrosino, A. (2001). Does research design affect study outcomes in criminal justice? *Annals of the American academy, 578*, s. 50–70.
West, S. G. & Thoemmes, F. (2010). Campbell's and Rubin's perspectives on causal inference. *Psychological methods, 15,* 18–37.
Wright, C. C. & Sim, J. (2003). Intention to treat approach to data from randomized controlled trials: A sensitivity analysis. *Journal of Clinical Epidemiology, 56*, 533–842.
Zwi, M., Jones, H., Thorgaard, C., York, A. & Dennis, J. (2012). Parent training interventions for Attention Deficit Hyperactivity Disorder (ADHD) in children aged 5 to 18 years. *Campbell Systematic Reviews, 2.* DOI: 10.4073/csr.2012.2

Kapitel 2 | Knut Sundell

Kort historik över experimentella utvärderingar

Experimentell forskning som vi förstår den i dag har existerat i drygt ett sekel. Det finns dock exempel där enskilda individer tillämpat vetenskapligt tänkande och experimenterande för att öka kunskapen om interventioners effekter.

År 1601 visade en engelsk kapten att en liten dos citronsaft dagligen förhindrar skörbjugg, en sjukdom som vid den här tiden tog många sjömäns liv. Det gjordes genom att kaptenen lät besättningen på ett skepp få tre skedar citronsaft per dag medan besättningen på två andra skepp inte fick det. Alla sjömän på försöksbåten överlevde medan 40 procent på kontrollskeppen avled i skörbjugg (Berwick, 2003). Denna kunskap fick dock ingen spridning. År 1747 gjorde James Lind, skeppsläkare på HM Bark Salisbury, ett experiment för att utveckla ett botemedel mot skörbjugg (Tansella, 2002). Lind valde ut 12 sjömän som led av skörbjugg och delade in dem i sex par. Varje par fick en speciell behandling under sex dagar. Samtliga behandlingar hade vid något tillfälle betraktats som effektiva:

- en liter cider per dag
- tjugofem droppar av vitriol (svavelsyra) tre gånger per dag på fastande mage
- en halv liter havsvatten varje dag
- en blandning av vitlök, senap och pepparrot i en klump stor som en muskotnöt

- två skedar ättika tre gånger om dagen
- två apelsiner och en citron varje dag.

De sjömän som fått citrusfrukter återhämtade sig dramatiskt. En av dem återvände till sin tjänst efter sex dagar och den andre blev sjukskötare för övriga sjuka. De andra sjömännen upplevde en viss förbättring, men ingenting var jämförbart med behandlingen med citrusfrukter.

Teorin utvecklas

Drygt hundra senare, mellan 1877 och 1883, utvecklade Charles Peirce, en amerikansk filosof, matematiker och forskare, en teori om statistisk slutledning. I den betonade han betydelsen av randomisering för att kunna dra säkra slutsatser. Peirce genomförde också ett experiment där han fördelade personer slumpmässigt till olika betingelser för att utvärdera deras förmåga att skilja mellan olika vikter. Han introducerade också maskering (eng. *blinding*) i experimentet, det vill säga att man hemlighåller vilka deltagare som får den ena eller den andra interventionen. Helst ska gruppindelningen hållas hemlig för såväl försökspersoner, försöksledare som för dem som värderar effekten för att förhindra att förväntningar påverkar behandlingsresultatet. Peirce inspirerade andra forskare inom psykologi och pedagogik som använde randomiserade laboratorieexperiment under slutet av 1800-talet.

År 1901 konstaterade de amerikanska pedagogerna Edward Thorndike och Robert Woodworth att de behövde en kontrollgrupp i sin forskning om att förbättra skolbarns psykiska funktioner. Bakgrunden var att forskare som studerade pedagogisk psykologi upplevde att den befintliga metodarsenalen inte var tillräcklig för att utesluta rimliga alternativa hypoteser. År 1908 var Thorndike och Woodworth bland de första att använda en utvärderingsdesign som bestod av förmätning, intervention och eftermätning i försöksgruppen och förmätning, ingen intervention och eftermätning i kontrollgruppen.

Ronald Fisher är en av 1900-talets mest inflytelserika statistiker. Han började 1919 arbeta på jordbruksforskningsstationen Rothamsted Experimental Station i Hertfordshire, England. Där studerade han bland annat hur olika gödningsmedel påverkade skördarna. Under de följande sju åren kom han att revolutionera experimentell metodik och utveckla tekniken för att statistiskt tolka experimentella resultat, bland annat genom variansanalys. Han är upphovsman till den design som kallas split-plot och som döpts efter att Fisher delade (eng. *split*) upp åkrar i olika delar (”plots”) och testade olika gödningsmedel på olika åkerlappar. År 1925 kulminerade hans arbete i boken *Statistical Methods for Research Workers* som blev ett standardverk för forskare från många olika discipliner. Tio år senare, 1935, publicerade han ett annat standardverk, *The Design of Experiments*.

Samma år startade *Cambridge-Somerville Youth Study* med pojkar från slummen i Boston, USA. Syftet var att förebygga kriminalitet bland pojkarna. Pojkarna lottades antingen till en grupp som fick stöd av socialarbetare under uppväxten eller till en grupp som inte fick ett sådant stöd. Varje barn i behandlingsgruppen tilldelades sedan en socialarbetare som försökte bygga upp en relation till pojken och som skulle vara till stöd för honom och hans familj under uppväxten. Familjen fick också hjälp med pojkens skolgång, hälsa och fritidsintressen. Denna studie om socialt arbete har blivit berömd för att man fortsatt att följa undersökningsgruppen i trettio år. När studien slutfördes 1945 visade resultaten att det gått sämre för interventionsgruppen (McCord, 1978). Nya analyser av det ursprungliga materialet antyder att resultaten kan förklaras med negativa gruppeffekter (McCord, 2003). De som klarade sig sämre över tid hade oftare deltagit i gruppbaserade aktiviteter inom projektet, till exempel sommarläger. Eftersom pojkarna tillhörde en riskgrupp med erfarenhet av kriminalitet kan gruppsamvaron ha ökat risken för antisocial inlärning (Dodge, Dishion & Lansford, 2006). Man brukar tala om en ”smittoeffekt”.

Även sociologer intresserade sig för experimentell forskning, något som stimulerades av den då pågående depressionen och önskan

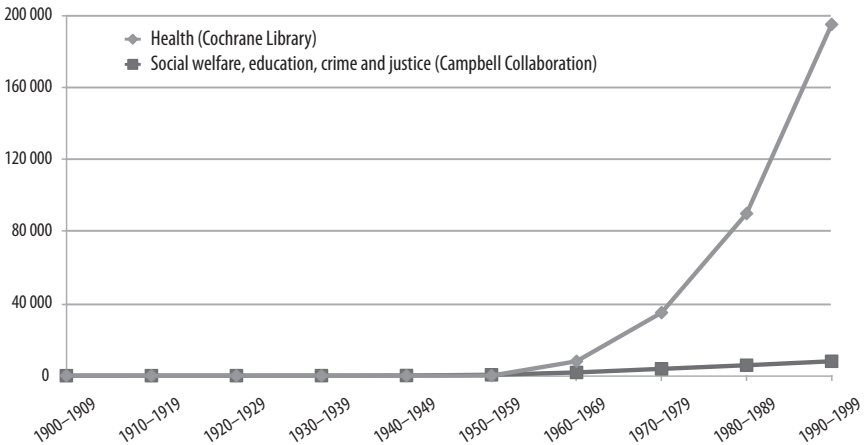
att finna effektiva sätt för att förbättra människors liv. Exempelvis hade minst 26 universitet i USA kurser i experimentell sociologi år 1931. År 1945 publicerade Ernest Greenwood boken *Experimental sociology* där han utvecklade argument för att använda experimentella metoder för studier av sociala fenomen. I den definierade han experiment som ”bevis om en kausal hypotes genom att studera två kontrollerade kontrasterande situationer”. Han rekommenderade användningen av fallstudier som förstudie till experimentella utvärderingar och betonade randomisering som ett överlägset sätt att säkra likvärdiga studiegrupper.

Experimentell forskning inom medicin

Det första kända publicerade randomiserat kontrollerade experimentet inom medicin var engelsmännen Austin Bradford Hills och Richard Dolls studie av tuberkulosbehandling från 1948. Vid det laget var experimentell forskning etablerad sedan flera decennier inom pedagogik, psykologi och sociologi. Därför är det paradoxalt att medicin totalt dominerade experimentell klinisk forskning vid millennieskiftet (figur 2:1). Även om placeboeffekten varit känd bland läkare sedan länge var det först på 1950-talet som forskare integrerade den i forskningsdesigner. År 1954 introducerade läkaren Henry Beecher dubbelblinda, placebokontrollerade kliniska försök. Beecher hade under andra världskriget observerat att när det saknades morfin och sjukvårdspersonal i desperation gav skadade soldater injektioner med destillerat vatten så lindrade det soldaternas smärtor. Inom psykiatri kom den första publicerade randomiserat kontrollerade studien 1955 när Davies och Shepherd använde en dubbelblind design där antipsykotiskt läkemedel jämfördes med placebo.

”The golden age of evaluation”

1960- och 70-talen har beskrivits som ”the golden age” för utvärdering i USA. En rad randomiserade fältstudier genomfördes om



Figur 2:1. Antal RCT från hälso- och sjukvård respektive socialtjänst, pedagogik och kriminalvård (källa: Shepherd, 2006).

bland annat kompensatorisk förskoleverksamhet för socialt utsatta barn (”Head-start”), försörjningsstöd, bostadsstöd, rehabiliteringsprogram för före detta fångar och villkorlig frigivning av fångar. Resultaten var emellertid nedslående. Få av de utvärderade interventionerna hade bättre effekter än det de jämförts med. Det medförde att forskare började studera varför interventioner fungerar, inte endast om de fungerar. En annan insikt var betydelsen av att implementera interventioner på rätt sätt, med hög behandlingstrohet.

Men allt går inte att studera med hjälp av randomiserade experiment. År 1966 publicerade Donald Campbell och Julian Stanley boken *Experimental and Quasi-Experimental Designs for Research*. I den argumenterar författarna för användningen av många alternativa designer för utvärdering där var och en har sina för- och nackdelar. Ett drygt decennium senare, 1979, publicerade Campbell tillsammans med Thomas Cook *Quasi-Experimentation: Design and Analysis Issues for Field Settings,* som kom att bli lärobok för många generationer forskare. En ny och utvidgad version av boken gavs ut 2002, nu med William Shadish som förstaförfattare, *Experimental and quasi-experimental designs for generalized causal inference*.

Health Technology Assessment

Hälso- och sjukvård har en lång historia av att använda någon form av evidens som underlag för politiska beslut. Det första dokumenterade svenska exemplet om att värdera sjukvård är från 1663 då *Collegium Medicorum* etablerades för att bland annat skilja kvacksalveri från medicin (Jonsson, 2009). Drygt tre århundraden senare, 1987, inrättades Statens beredning för medicinsk utvärdering (SBU).[1] SBU och andra organisationer för Health Technology Assessment (HTA) spelar en viktig roll i utvecklingen av kriterier för vilka effektutvärderingar som kan tillskrivas tillräcklig trovärdighet för att ingå i kunskapsunderlag om medicinska interventioners effekter. Den forskare som i dag vill att hans eller hennes effektutvärderingar ska ingå i underlaget för att värdera interventioners effekter (och därmed ha inflytande över policy) måste således anpassa sig efter de kriterier som ställts upp.

Den internationellt mest kända HTA-organisationen är Cochrane Collaboration, som är en internationell, oberoende, icke-kommersiell organisation med över 28 000 bidragsgivare från mer än 100 länder. Cochrane Collaboration bildades 1993 för att syntetisera kunskap från hälso- och sjukvård via systematiska översikter av forskningslitteraturen.

Systematiska översikter belyser även en annan fråga. En enskild effektutvärdering ger i allmänhet sämre information om resultatens generaliserbarhet eftersom det oftast bara ingår en eller några få arbetsplatser. En systematisk översikt som består av flera utvärderingar av samma intervention men genomfört av olika forskare från olika platser ger bättre möjlighet att tolka resultatens externa validitet. En central fråga är att avgöra vilka studier som ska ingå i översikterna och vilka som ska exkluderas. I Cochranes handbok[2] beskrivs bland annat kriterier för detta.

---

1 Det är så vitt känt är världens äldsta existerande institut för syntetisering av kunskap.

2 http://www.cochrane-handbook.org/

År 2000 startade Campbell Collaboration med systematiska översikter inom kriminalvård, socialt arbete och pedagogik. Det följdes år 2001 av Nordiska Campbell Collaboration med säte i Köpenhamn (år 2008 omvandlades det till SFI Campbell). I Sverige inrättades Institutet för utveckling av metoder i socialt arbete (IMS) 2004 som kommit att arbeta med systematiska översikter inom området socialt arbete. Från och med 2010 ingår IMS i Socialstyrelsens linjeorganisation, där översiktsarbete nu bedrivs inom enheten för kunskapsöversikter.

Standardisering av experimentella studier

Samma år som Cochrane Collaboration bildades, 1993, träffades en grupp forskare och redaktörer för vetenskapliga tidskrifter för att diskutera hur man skulle kunna förbättra rapporteringen av experimentella studier. Mötet resulterade i *Standardized Reporting of Trials* (SORT) som bestod av en checklista och ett flödesdiagram som forskare uppmanades att använda för sin rapportering av hur de genomfört randomiserat kontrollerade studier. Samtidigt hade en annan grupp experter samlats för att ställa samman rekommendationer för hur randomiserat kontrollerade studier inom biomedicinsk forskning skulle göras. Båda dessa grupper samordnade sitt arbete och publicerade *Consolidated Standards of Reporting Trials* (CONSORT) 1996. Därefter har två uppdaterade versioner av CONSORT publicerats, dels 2001, dels 2010 (Moher, Schulz & Altman, 2001; Moher, Hopewell, Schulz, Montori, Gøtzsche, Devereaux m.fl., 2010). Flera vetenskapliga tidsskrifter kräver i dag att artiklar använder CONSORT.

Utöver CONSORT finns även riktlinjer för hur icke-randomiserade utvärderingar ska rapporteras, *The Transparent Reporting of Evaluations with Non-randomized Designs* (TREND) samt hur observationsstudier inom epidemiologi ska rapporteras, *The Strengthening the Reporting of Observational Studies in Epidemiology* (STROBE).

År 2000 formerades The Grading of Recommendations Assess-

ment, Development and Evaluation (GRADE) Working Group som ett informellt samarbete för personer som såg behovet av att utveckla graderingssystem för interventioner inom hälso- och sjukvård. Ett växande problem var att olika organisationer använde olika system för att gradera trovärdigheten för utvärderingar. Samma forskningsunderlag kunde generera vitt skilda rekommendationer, allt från att en intervention klassificerades som ”rekommenderas starkt” eller ”lovande” till ”okänd effekt”. GRADE har utvecklat ett kategoriseringssystem som väger samman information om design, dos-respons-relation, storlek på effekter med mera. Även dessa kriterier ger vägledning till forskare om vad som är väsentligt för att forskningsresultaten ska få inflytande.

Ett av de senaste bidragen till att öka kvaliteten i effektutvärderingar publicerades 2005 av en grupp forskare med anknytning till The Society for Prevention Research (SPR) (Flay m.fl., 2005). På uppdrag av SPR formulerade forskarna förslag till normer för utvärderingar av förebyggande interventioner. Förslaget skiljer mellan modellutvärderingar (eng. *efficacy trials*) och utvärderingar av ordinarie verksamhet (eng. *effectiveness trials*). För att en intervention ska betraktas som effektiv i modellutvärderingar så måste det finnas minst två experimentellt kontrollerade studier där (1) interventionen är beskriven så att andra kan replikera den, (2) undersökningsgruppen är tydligt definierad, (3) utvärderingarna använt psykometriska tillförlitliga mått och datainsamlingsprocedurer, (4) det finns statistiskt säkerställda skillnader (och inga negativa effekter), (5) skillnaderna är konsistenta mellan olika utvärderingar och (6) det finns minst en långsiktig uppföljning. För att en intervention i en ordinarie verksamhet ska betraktas som effektiv krävs dessutom (7) att interventionen är effektiv i ordinarie verksamhet, (8) att det finns en manual som beskriver hur interventionen ska implementeras, (9) att det finns en tydlig teori om orsaksmekanismer, (10) resultat som visar på den praktiska betydelsen av intervention (utöver att den är statistiskt skillnad) samt (11) att utvärderingen tydligt demonstrerat för vilka resultaten kan generaliseras. Utöver detta

anger SPR också kriterier för vad som gäller för interventioner som är kvalitetssäkrade för att kunna spridas brett.

Utvecklingen fortsätter

Som framgår av denna historik har experimentell forskning utvecklats under drygt hundra år. Det senaste decenniets bidrag handlar om att en randomiserat kontrollerad studie inte räcker för att säkra tillförlitliga resultat. Exempelvis betonas i dag att alla som randomiserats till en utvärdering måste ingå i analyserna, på engelska kallat *Intention to treat analysis*, *ITT* (Wright & Sim, 2003), i motsats till tidigare då analyser oftast endast baserades på dem som fullföljt en behandling, på engelska kallat *Treatment on the treated*, *TOT*. ITT ställer krav på att personer som hoppat av en studie tilldelas ett värde genom statistik eller andra tekniker. Betydelsen av att dokumentera interventionens kvalitet framhålls också på ett annat sätt än tidigare för att undvika att dåligt implementerade interventioner medför att en i grunden effektiv intervention förkastas. Ett tredje exempel på förändrat synsätt är en ökad förståelse för att undersökningspersoner ibland kan påverka varandra, exempelvis elever som går i samma klass eller skola, och att detta beroende behöver hanteras vetenskapligt.

Ovanstående innebär att äldre effektutvärderingar successivt blir mindre betydelsefulla som underlag för att värdera interventioners effekter. Rent allmänt kan man säga att utvärderingar som är gjorda på 2000-talet håller högre kvalitet än de som är publicerade tidigare.

Fördjupningslitteratur

Oakley, A. (1998). Experimentation and social interventions: a forgotten but important history. *Education and debate, 317*, 1239–1242.

Salsburg, D. (2001). *The lady tasting tea. How statistics revolutionized science in the twentieth century*. New York: Owl books.

Tansella, M. (2002). The scientific evaluation of mental health treatments: an historical perspective. *Evidence Based Mental Health, 5,* 4–5.

Referenser

Berwick, D. M. (2003). Disseminating innovations in health care. *JAMA, 289*, 1969–1975.

Dodge, K. A., Dishion, T. J. & Lansford, J. E. (red.) (2006). *Deviant Peer Influences in Programs for Youth Problems and Solutions.* New York: Guilford Press.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., Mościcki, E. K., Schinke, S., Valentine, J.C. & Ji, P. (2005). Standards of evidence. Criteria for efficacy, effectiveness and dissemination. *Preventions science, 6,* 151–175.

Jonsson, E. (2009). History of health technology assessment in Sweden. *International Journal of Technology Assessment in Health Care, 25,* 42–52.

McCord, J. (1978). A thirty-year follow-up of treatment effects. *American Psychologist* 33 (3): 284–89.

Moher, D., Schulz, K.F. & Altman, D.G. (2001). The CONSORT statement: revised recommendations for improving the quality of reports of parallell group randomized trials. *BMC Medical Research Methodology, 1,* 2.

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., m.fl. (2010). CONSORT 2010 Explanation and elaboration: updated guidelines for reporting parallel group randomised trials. Research Methods & Reporting. *BMJ, 340*: c869 (doi: 10.1136/bmj.c869).

Oakley, A. (1998). Experimentation and social interventions: a forgotten but important history. *Education and debate, 317*, 1239–1242.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.

Shepherd, J. (2006). The production and management of evidence for public service reform. *Evidence & policy, 3,* 231–251.

Tansella, M. (2002). The scientific evaluation of mental health treatments: an historical perspective. *Evidence Based Mental Health,* 5, 4–5.

Wright, C. C. & Sim, J. (2003). Intention to treat approach to data from randomized controlled trials: A sensitivity analysis. *Journal of Clinical Epidemiology, 56*, 533–842.

Kapitel 3 | Ann-Charlotte Smedler

Forskningsetik

Samhället i stort har ett legitimt intresse av ny kunskap, men denna kunskap får inte vinnas på enskilda människors bekostnad. Forskning på människor måste utföras med respekt för människovärdet och får inte stå i konflikt med grundläggande rättigheter och friheter. Människors välfärd ska alltid ges företräde framför vetenskapens behov. Det kan låta som självklarheter, men dessa etiska principer har en ganska kort historia. Det var först efter andra världskrigets slut, i samband med Nürnbergrättegångarna, som den första forskningsetiska koden formulerades. I Nürnberg framkom det att medicinska forskare hade bedrivit hänsynslösa humanexperiment i koncentrationslägren. Den experimentella kontrollen var total och mänskliga rättigheter obefintliga. Forskningen hade inneburit ytterligare svåra kränkningar och lidanden för de internerade.

Situationen i Tredje riket var en ytterlighet, men inte så unik som vi kanske vill tro. Även i Sverige förekom det att människor tvingades delta i vetenskapliga studier som innebar allvarliga kränkningar. Ett slående exempel är den odontologiska studie som under åren 1946–1952 genomfördes på människor med förståndshandikapp boende på institutionen Vipeholm strax utanför Lund. På denna totala institution kunde man experimentellt kontrollera de boendes kost och sockerintag och studera hur dessa faktorer påverkade munhälsa och tandstatus. Studien innebar många provtagningar och hade ock-

så långsiktiga hälsoeffekter. Vipeholmsundersökningarna genomfördes på människor med svåra funktionshinder, som dessutom befann sig i en total beroendesituation. Undersökningarna gjordes på uttryckligt uppdrag av regering och riksdag och under överinseende av Medicinalstyrelsen (föregångare till Socialstyrelsen), och ingen ansvarig tycktes ha några som helst etiska betänkligheter. Tvärtom framhölls det som en stor fördel att institutionen erbjöd ett ”omfattande folkmaterial som helt kunde hållas under kontroll under avsevärd tid” (Svensk Tandläkaretidskrift, 1948, s. 2).

Studien gav ny kunskap om vikten av att begränsa sockerintag och småätande samt iaktta god munhygien för att förebygga kariesangrepp och andra ohälsotillstånd. Denna kunskap gav upphov till den goda vanan att låta barn få ”lördagsgodis” i stället för dagliga sötsaker och att uppmuntra ordentlig tandborstning. Vipeholmsstudierna har alltså haft avgörande betydelse för att senare generationers barn har betydligt bättre tandstatus än personer födda på 1950-talet eller tidigare. Forskningsresultaten har alltså haft en direkt och positiv påverkan på folkhälsan. Det rättfärdigar dock inte de grova fysiska och psykiska kränkningar som de boende på Vipeholm utsattes för. Bo Petersson, professor i praktisk filosofi i Linköping, gjorde omkring 1990 en ingående granskning och forskningsetisk analys av Vipeholmsundersökningarna. Tack vare hans rapport har detta mörka men illustrativa exempel sent omsider blivit känt och diskuterat (Petersson, 1994).

Ironiskt nog pågick Vipeholmsprojektet samtidigt som Nürnbergrättegångarna. I Nürnberg uppenbarades att det hade begåtts allvarliga övergrepp med vetenskapliga förtecken. Visserligen var förhållandena i Tredje riket extrema, men de kastade ljus på att det finns ett potentiellt etiskt dilemma inom all humanforskning, mellan vetenskapliga krav å ena sidan och individens rättigheter å den andra. Med Nürnbergrättegångarna tydliggjordes behovet av en normativ etik inom forskningens område. Den normativa etikens uppgift är att klargöra hur etiska problem bör hanteras och att besvara frågan vad som är den rätta handlingen (Kagan, 1998).

Forskningsetik i den mening som begreppet kommer att användas här handlar alltså främst om hur man bör handla så att människor inte kommer till skada genom forskningen.

Forskningsetikens framväxt

Den första Nürnbergkodexen låg klar 1947, och med den slogs kravet på *informerat samtycke* fast. Det är ett viktigt uttryck för en bredare *autonomiprincip*, individens rätt att bestämma över sig själv och sin medverkan. Vidare uttrycktes att forskningen ska ha goda konsekvenser, ett uttryck för *göra gott-principen* (eng. *beneficence*). Riskerna för försökspersoner ska minimeras, och om det trots vidtagna försiktighetsåtgärder verkar troligt att en undersökningsdeltagare kan komma att påverkas negativt ska forskaren avbryta försöket. Detta är ett uttryck för *principen att inte skada* (eng. *non-maleficence*). Ytterligare en hörnsten i den normativa medicinska etiken är *rättviseprincipen*, som utgår från alla människors lika värde och formulerar krav på icke-diskriminering samt att särskilda hänsyn ska tas till människor som har försvagad förmåga att hävda sina egna intressen (Beauchamp & Childress, 2009). I Nürnbergkodens anda följde snart andra viktiga deklarationer, bland annat FN:s allmänna förklaring om mänskliga rättigheter från 1948 och Europakonventionen till skydd för de mänskliga rättigheterna och grundläggande friheterna från 1950. De etiska principer som formulerats för att tillgodose individskyddet visavi forskningen handlar alltså om grundläggande mänskliga rättigheter.

Nürnbergkoden formulerades i första hand för medicinsk forskning, och alltsedan dess har den normativa etiken till skydd för undersökningsdeltagarna i första hand haft sikte på medicinsk vetenskap. Det finns goda skäl till detta. Medicinarnas arbetsfält rör frågor om liv och död, och människor som är under vård befinner sig i ett beroendeförhållande och uppenbart underläge. Tydliga etiska principer, så att undersökningsdeltagarnas rättigheter och intressen skyddas, har därför en särskild aktualitet för medicinsk forskning.

Här bör även nämnas den så kallade Helsingforsdeklarationen, som togs fram av *World Medical Association* år 1964 (senast uppdaterad 2008). Med denna infördes bland annat en distinktion mellan terapeutisk och icke-terapeutisk forskning.

Specifika etiska riktlinjer för samhälls- och beteendevetenskaplig forskning utvecklades senare än den medicinska etiken. De bygger dock på samma grundläggande principer. I synnerhet vid interventionsforskning är det liknande överväganden som är aktuella, oavsett om det är en beteendevetenskaplig eller medicinsk behandling som är i fokus.

Vad är speciellt med interventionsstudier?

Interventionsstudier ställer särskilda etiska krav, eftersom studiedeltagarna är föremål för påverkan. Detta är en avgörande skillnad visavi forskning som enbart bygger på registrering av sakförhållanden, utan aktiv påverkan. Självklart kan även den senare typen av forskning inrymma etiska komplikationer. Sakförhållandena kan röra personliga förhållanden av mer eller mindre känslig natur, och forskarnas frågor kan upplevas som integritetskänsliga. I vissa fall kan de frågor som ställs via enkäter och intervjuer väcka minnen och känslor till liv, och på så vis oavsiktligt påverka forskningspersonen. Denna möjliga påverkan är givetvis något som forskaren bär ansvar för och måste hantera på bästa sätt, men den är inte forskningens syfte utan snarast en bieffekt. Ibland kan sådana bieffekter te sig så riskfyllda att man bör avstå från att genomföra undersökningen.

Men interventionsforskningens själva *syfte* är att påverka, och att avläsa om effekterna blir de avsedda. Vi förväntar oss alltså att en interventionsstudie ska få konsekvenser för forskningspersonerna. Avsikten är att effekterna ska vara goda och gynna människors hälsa och utveckling. Men att så verkligen är fallet kan vi inte veta förrän interventionen prövats i vetenskapliga studier. Vetenskapliga försök på människor är alltså nödvändiga för att utveckla interventionsmetoder som kan hjälpa människor. Men det innebär alltid

ett visst risktagande att pröva nya interventionsmetoder, eller att pröva etablerade metoder i nya sammanhang. Innan interventionen varit föremål för vetenskapligt studium vet vi ju inte om den har avsedd effekt.

För att avläsa effekterna av en intervention behöver också uppgifter om de deltagande personerna samlas in och förvaras hos forskarna. Om man ska kunna uttala sig om en interventions effekter måste deltagarna följas över tid. De data man är intresserad av innehåller oftast integritetskänsliga uppgifter om hälsa och personliga förhållanden. Interventionsstudier innebär att forskaren, ensam eller tillsammans med en vårdgivare, iklär sig ett dubbelt ansvar visavi forskningspersonerna. Forskaren har ansvar både för deras välbefinnande i samband med att de deltar i interventionen och för att den integritetskänsliga information som de bidragit med inte används för andra syften än forskning, att den skyddas av sekretess och att datasäkerheten är fullgod.

Nedan följer en närmare belysning av etiska principer och överväganden, så som de kan aktualiseras i samband med interventionsstudier. Följande teman kommer att diskuteras. De första tar sin utgångspunkt i var och en av de fyra forskningsetiska huvudprinciper som i dag är allmänt accepterade (Gillon, 2003): *principen att göra gott, principen att inte skada, autonomiprincipen* och *rättviseprincipen*. Frågor om *konfidentialitet* och *datasäkerhet* berörs helt kort, och därefter följer en kort diskussion om *jävsförhållanden* och *vetenskaplig redlighet*. Kapitlet avslutas med några basfakta och råd kring *forskningsetisk prövning*.

Principen att göra gott – beneficence

Själva syftet med en intervention (i den mening vi talar om den här) är att göra gott. Det gäller oavsett om interventionen är en behandling eller en preventiv metod. Rör det sig om behandling förväntas den minska deltagarnas symtom och lidanden, och i bästa fall undanröja dessa. Är det en förebyggande intervention förväntas den

förstärka skyddsfaktorer hos deltagarna, vilket i sin tur minskar risken för framtida ohälsa.

Dessa goda syften innebär att det kan ses som ett moraliskt imperativ att erbjuda effektiva interventioner. Men hur vet vi att interventionen verkligen är effektiv och gör gott? För att besvara den frågan behövs bra studier.

Ansvar gentemot kontrollgruppen

Ganska ofta hävdas det att det vore etiskt betänkligt att ha en kontrollgrupp som inte får tillgång till interventionen, och att det därför rentav vore oetiskt att designa en randomiserad kontrollerad studie (RCT). En sådan invändning kan ha fog för sig, om 1) de presumtiva studiedeltagarna är i akut behov av stöd och behandling och 2) preliminära resultat tyder på att den aktuella interventionen verkligen är effektivare än standardbehandling (eng. *treatment as ususal, TAU*) eller ingen behandling alls. Under förutsättning att *båda* dessa krav är uppfyllda kan det finnas etiska skäl att avstå från en RCT-design, som skulle lämna kontrollgruppen utan intervention. Om det ändå är etiskt försvarbart att fördröja interventionen kan de som randomiserats till kontrollgruppen få del av den senare och tills vidare stå på väntelista. Är även detta betänkligt återstår att optimera kontrollen via multipla baslinjemätningar eller en single-case-design.

Förment omsorg om kontrollgruppen får dock inte leda till att man avstår från att göra välkontrollerade studier. Ibland kan en stark övertygelse om interventionens förtjänster, snarare än vetenskapliga bevis, ligga i vägen för att den verkligen prövas – ofta med hänvisning till att det vore oetiskt att inte låta personerna i kontrollgruppen få del av den. I själva verket kan prematura slutsatser om en interventions effektivitet hindra att den blir föremål för nödvändig vetenskaplig prövning. Det om något vore etiskt betänkligt.

Hur mycket behöver vi veta beträffande en interventions potentiella värde för att det ska vara etiskt försvarbart att initiera en effektstudie? I interventionsforskning finns det etiska skäl att vara mycket försiktig med att pröva det som Karl Popper kallade ”djärva

hypoteser". Den omsorgsfulla forskaren förvissar sig i stället om att interventionen bygger på bästa tillgängliga kunskap för att åstadkomma ett positivt utfall, och att riskerna för eventuella skador till följd av interventionen i sig eller övriga forskningsinslag minimeras så långt det är möjligt. Det innebär bland annat att en intervention bör bli föremål för en eller flera modellutvärderingar (eng. *efficacy*) innan det är aktuellt med en utprövning av metoden i ordinarie verksamhet (eng. *effectiveness*).

I en modellutvärdering prövas interventionen under välkontrollerade och bästa möjliga förhållanden, vilket optimerar möjligheterna att fånga upp faktiska positiva effekter. En intervention som inte visat sig lovande i minst en modellutvärdering är troligen verkningslös och knappast etiskt försvarbar att pröva under mindre välkontrollerade och mer realistiska förhållanden, i en verksamhetsutvärdering (eng. *effectiveness*). Inom det sociala och beteendevetenskapliga fältet kan man diskutera var gränsen går mellan en studie som prövar metodens "efficacy" respektive "effectiveness". Det finns en glidande skala från prövningar med hög experimentell kontroll och strikta inklusions- och exklusionskriterier till studier som görs inom ramen för ordinarie verksamheter. Ett nytt behandlingsprogram kan vara uppbyggt av komponenter som visat sig effektiva i tidigare studier, och det kan då vara välmotiverat att låta en första prövning ha mer av effectiveness-karaktär.

"Bästa tillgängliga kunskap" begränsar sig inte nödvändigtvis till tidigare interventionsstudier och kännedom om aktuella risk- och friskfaktorer. Även kunskap om specifika sociokulturella omständigheter kan vara nödvändiga (jfr kapitel 7). Då en metod prövas i nya sammanhang och användningsområden bör det föregås av en allsidig analys av på vilket sätt förhållandena skiljer sig åt. Självklart bör man också göra pilotstudier, och även ha beredskap att hantera eventuella negativa reaktioner. Det förekommer dock att tillskyndare av en interventionsmetod drivs av en övertygelse att den är användbar i nya sammanhang och snabbt vill göra den tillgänglig inom nya användningsområden. Man kan då frestas att hoppa

över steg i den här processen. Ett aktuellt exempel på detta är den snabba implementeringen av ACT (Acceptance and commitment therapy), en KBT-baserad intervention som använder strategier för acceptans, medveten närvaro och beteendeförändring för att öka individens psykologiska flexibilitet och välbefinnande (Hayes, Luoma, Bond, Masuda & Lillis, 2006). Metoden har visat god effekt vid en rad olika tillstånd. Detta har lett till att den snabbt spritts, också till användningsområden för vilken den inte har någon vetenskaplig evidens.

Nyligen fick en ACT-studie avslag på etikansökan, just med hänvisning till att den tilltänkta interventionen var alltför oprövad. Forskargruppen ville pröva en telefonbaserad ACT-intervention för isolerade äldre människor med nedstämdhetssymtom. Ansökan avslogs eftersom det ännu inte finns några studier av ACT:s eventuella effekter för den aktuella åldersgruppen och problematiken. Man vet därför inte om metoden, ens om den är implementerad inom ramen för en terapeutisk kontakt, har förutsättningar att förbättra isolerade äldre människors välbefinnande. Att då gå direkt på en likaledes oprövad telefonadministrerad intervention bedömdes inte etiskt försvarbart.

Principen att göra gott manar oss att tänka en extra vända kring hur kontrollgruppen hanteras. Det vanliga är att kontrollgruppen får del av standardbehandling – vilket ibland innebär inga särskilda åtgärder alls. Av vetenskapliga och metodologiska hänsyn, om än inte av etiska, finns det goda skäl att särskilt dokumentera vad standardbehandling innebär. Däremot är det knappast etiskt försvarbart att försöka begränsa kontrollgruppens tillgång till standardbehandling.

Exkluderade deltagare

I en behandlingsstudie vill vi vanligen undersöka om interventionen har effekt på en viss typ av problematik. I synnerhet om det rör sig om en modellutvärdering kan det vara aktuellt att renodla effektfrågeställningen. Man vill då rekrytera studiedeltagare som har just denna problematik, men däremot inga ytterligare svårigheter av all-

varlig art. Det är också vanligt att man vänder sig till en viss åldersgrupp eller har andra demografiska urvalskriterier för de tilltänkta deltagarna. Därutöver kan det finnas specifika exklusionskriterier, till exempel att deltagarna inte ska ha pågående behandling av annat slag. Det är ofta vetenskapligt välmotiverat med ganska snäva urvalskriterier; man minskar då andelen varians som designen inte har förutsättningar att förklara och har bättre möjligheter att uttala sig om interventionens effekter för en viss problematik. För att få den väldefinierade undersökningsgrupp man önskar behöver man göra en noggrann screening bland presumtiva undersökningsdeltagare. Bara de som uppfyller inklusionskriterierna och inga eventuella exklusionskriterier blir i slutändan inbjudna att delta. Men det kan medföra en etisk komplikation att avvisa intresserade som kanske också verkligen är i behov av behandling. Vilket ansvar har forskaren gentemot dem?

På den frågan finns inget generellt svar. Men man bör i förväg tänka igenom hela proceduren, så att risken för att en exkluderad person ska känna sig illa bemött minimeras. Här är förhandsinformationen viktig; av den första information som används vid rekryteringen ska det framgå att det kommer att göras en screening som alla presumtiva deltagare kommer att bli föremål för och vad som är syftet med den.

Beroende på interventionens karaktär kan man överväga om man ska ge någon särskild feedback eller särskilda råd till dem som inte uppfyller inklusionskriterierna. Det är inget självklart inslag, men kan vara motiverat om det rör sig om en intervention som kan intressera personer med tydliga behandlingsbehov. Man bör på förhand ha tagit ställning till om och i så fall hur samt under vilka förutsättningar man kommer att ta ansvar för vidare remittering. Görs interventionsstudien inom ramen för till exempel socialtjänsten eller hälso- och sjukvården kan man förvissa sig om att personen har tillgång åtminstone till standardbehandling.

Principen att inte skada – non-maleficence

Principen att göra gott kräver att studien har en sund design och håller god vetenskaplig kvalitet. Det är en förutsättning för att effektfrågan ska kunna besvaras. Göra gott-principen förutsätter också att forskarna har kompetens att genomföra projektet i sin helhet på ett professionellt sätt och att skydda forskningsdeltagarna från obehag och otillbörligt intrång i den personliga sfären. Ändå är det ofrånkomligt att forskningsprocessen innebär ett moment av risktagande. Enligt internationella överenskommelser får forskning på människor genomföras enbart om det inte finns något alternativt och mindre riskabelt sätt att få motsvarande kunskap. En vägning mellan risken för skada och möjligheten att göra gott är central i forskningsetisk granskning. Den potentiella nyttan med forskningen måste vara tveklöst större än riskerna. Om det föreligger risk för allvarligt obehag eller skada finns det etiska skäl att avstå från att genomföra studien, även om den i andra avseenden har ett stort potentiellt värde. Det finns alltså en punkt där en vägning mellan möjliga positiva och negativa konsekvenser inte längre är relevant. Vissa risker är helt enkelt oacceptabla. Även om de berörda försökspersonerna har gett sitt informerade samtycke går det inte att etiskt försvara studien.

Principen att inte skada manar alltså till försiktighet.[1] Det gäller i särskilt hög grad om forskningsdeltagarna inte själva är beslutskompetenta och inte kan förmodas ha nytta av forskningen för egen del. I sådana situationer ställs särskilt höga krav på att forskningen ska vara angelägen, att ingen alternativ metodik kan ge motsvarande kunskap och att man aktivt ska bevaka att forskningspersonerna utsätts för minimal risk och belastning. Med det menas att obehag, om det uppstår, endast får vara lindrigt och övergående.

1 Ibland hör man också att uttrycket ”försiktighetsprincipen” används, oftast i det närmaste synonymt med principen att inte skada.

Rapporteringssystem för negativa utfall

I interventionsstudier är givetvis förhoppningen att de som deltar i interventionen själva ska dra nytta den. Det betyder att principen att inte skada har högsta prioritet. Även om man har planerat studien omsorgsfullt kan man inte utesluta att vissa forskningsdeltagare reagerar negativt på interventionen eller på andra delar av forskningsprocessen. Ur såväl vetenskaplig som etisk synvinkel är det att rekommendera att forskningsprotokollet innehåller ett rapporteringssystem för negativa effekter. Det hör tyvärr mer till undantagen än till reglerna att beteende- och socialvetenskapliga interventionsstudier har ett sådant system inbyggt. Om negativa effekter skulle rapporteras på ett mer systematiskt sätt skulle specifika riskfaktorer kunna identifieras snabbare – vilket skulle ha uppenbara etiska fördelar, men även ge vetenskapliga vinster.

De negativa utfall som finns rapporterade i litteraturen tycks oftast ha uppdagats då forskarna post hoc har närstuderat variationen i effekter och funnit en systematik som man dittills varit omedveten om. Ett känt exempel är Dishions och medarbetares rapporter från 1990-talet, som redovisade att det kan vara kontraindicerat att sammanföra utagerande tonåringar i en gruppintervention (Dishion & Andrews, 1995). I själva verket ökade både symtomen på psykisk ohälsa och det allmänt normbrytande beteendet hos de ungdomar som deltagit i gruppinterventionen. Att i stället engagera ungdomarnas föräldrar i interventionen visade sig ha en liten men gynnsam effekt, vilket även andra studier visat (Warren, Mober & McDonald, 2006). Efter att ha fått upp ögonen för denna möjliga negativa interventionseffekt genomförde Dishion och medarbetare flera studier som visade att det finns en klar risk för långsiktig, negativ socialisering i grupper som inkluderar barn och ungdomar med utagerande symtom (Dishion, Poulin & Burraston, 2001; Dishion, McCord & Poulin, 1999). Det här fenomenet har beskrivits med termer som ”social smitta” och ”inskolning i en avvikarroll”. Det har visat sig att även tidigare problemfria ungdomar kan påverkas i negativ riktning (Cavell & Hughes, 2000; Palinkas, Atkins, Miller &

Ferreira, 1996; Biederman, Faraone, Monuteaux & Feighner, 2000).

Enligt den första Nürnbergdeklarationen och efterföljande etiska riktlinjer ska forskaren omedelbart avbryta försöket om en undersökningsdeltagare kan komma att påverkas negativt. Det förutsätter egentligen att en systematisk bevakning av negativa reaktioner och effekter finns inbyggd i undersökningsprotokollet – något som långt ifrån alltid är fallet. Man kan bygga in en rutin för att särskilt uppmärksamma och följa upp tecken på försämringar hos individuella deltagare. Återkommande korta utvärderingar med deltagarna kan också fylla sin funktion. Vid interventionsstudier måste man dock räkna med att de negativa effekterna kan visa sig på lite sikt, och inte nödvändigtvis i direktkontakt mellan deltagare och ansvarig forskare. Både då studien pågår och senare är det viktigt att deltagarna har kontaktuppgifter till ansvarig forskare och känner sig uppmuntrade att höra av sig om det uppstår några frågor.

Redan då en interventionsstudie planeras bör man tänka igenom hur negativa utfall ska hanteras. Utöver ett systematiskt rapporteringssystem kan det vara klokt att ha en backup, så att den som reagerat negativt eller avbryter behandlingen har möjlighet att få en individuellt anpassad kontakt med en behandlare.

Autonomi – det informerade samtyckets princip

Autonomiprincipen handlar om personens rätt att bestämma över sig själv. Respekten för den personliga integriteten, det vill säga människors rätt att bestämma över en personlig sfär, går att härleda till denna princip. Att ge sig in i denna personliga sfär utan att ha fått godkännande innebär en kränkning av den personliga integriteten (Hermerén, 1996). Det kan handla om att ta del av känsliga personuppgifter utan att ha fått de berördas tillstånd, eller att inkludera personer i till exempel en interventionsstudie utan att de informerats och beretts tillfälle att samtycka eller avböja deltagande.

Information till studiedeltagare

Det är en forskningsetisk huvudregel att studiedeltagarna ska ha gett sitt *informerade samtycke* till att medverka i en studie. Informerat samtycke används ofta som ett sammanhållet begrepp, men information och samtycke behöver uppmärksammas var för sig. En förutsättning för att kunna ge ett i egentlig mening *informerat* samtycke är att man har fått adekvat information. Har den varit bristfällig är det oklart vad det är man egentligen samtycker till. Hur ska man då informera de tilltänkta studiedeltagarna?

Kring detta finns det olika råd och riktlinjer, en del av dem mycket detaljerade. Det går dock att urskilja ett mindre antal gemensamma huvudpunkter, som i praktiken kan fungera vägledande vid all forskning som avser människor. I den svenska etikprövningslagen, som trädde i kraft 2004, uttrycks informationskravet[2] på ett sätt som överensstämmer väl med de flesta andra direktiv:

> Forskningspersonen skall informeras om:
> – den övergripande planen för forskningen,
> – syftet med forskningen,
> – de metoder som kommer att användas,
> – de följder och risker som forskningen kan medföra,
> – vem som är forskningshuvudman,
> – att deltagande i forskningen är frivilligt och
> – forskningspersonens rätt att när som helst avbryta sin medverkan.[3]

De sista punkterna är viktiga men enkla sakupplysningar. Men de fyra första fyra punkterna kräver fortfarande sin uttolkning. Hur detaljerad bör jag vara då jag beskriver forskningens syfte, plan och

2 SFS (2003:460). Lagen om etikprövning av forskning som avser människor (EPL) 16 §. www.riksdagen.se

3 Enligt 19 § EPL får data som dessförinnan har hämtats in användas i forskningen. Det innebär exempelvis att forskaren kan inkludera den f.d. deltagarens baselinedata i en ”intent-to-treat analysis”. I vissa sammanhang och beroende på hur förhandsinformation och samtycke har formulerats kan det vara mer korrekt att stryka deltagarens uppgifter helt och hållet.

metodik? Och hur informerar jag om forskningens följder och eventuella risker, utan att vare sig skönmåla eller onödigtvis skrämma någon? I det svenska före detta Humanistisk-samhällsvetenskapliga forskningsrådets etiska principer sammanfattas essensen i informationskravet på följande sätt:

> Informationen skall omfatta alla de inslag i den aktuella undersökningen som rimligen kan tänkas påverka deras villighet att delta.[4]

I denna formulering ryms en uppmaning till empatisk reflektion: Vad är viktigt ur den tilltänkta deltagarens perspektiv? Självklart behöver hon eller han få information om sin uppgift i projektet och vilka villkor som gäller för deltagande. Det är viktigt att veta att deltagandet är frivilligt och att man har rätt att när som helst avbryta sin medverkan. Både sådant som kan öka och sådant som kan minska intresset av att delta är relevant att informera om. Vid interventionsstudier kan förväntningar om positiv effekt vara ett starkt motiv till att vilja delta. I förhandsinformationen är det därför viktigt att understryka att studien görs för att undersöka om det verkligen finns en sådan effekt. Man ska inte undanhålla eventuella negativa inslag, och ingen ska behöva känna sig obehagligt överraskad av vad det visade sig innebära att delta i forskningen. Men man ska inte heller försumma att informera om positiva aspekter, till exempel studiens långsiktiga syfte.

Om man randomiserar deltagarna till olika aktiva betingelser, till exempel till intervention och uppmärksamhetskontroll, bör man undvika att avslöja sina hypoteser om vad som har effekt. En preciserad information om forskningshypoteserna skulle ju undergräva själva designen. I stället bör man i efterskott och så snart det är möjligt ge kompletterande information och förklara tillvägagångssättet.

Informationen ska ges på ett enkelt och tydligt språk och inte

4 Forskningsetiska principer inom humanistisk-samhällsvetenskaplig forskning. (1990). Vetenskapsrådet. www.codex.vr.se/texts/HSFR.pdf, s. 7.

innehålla ord som kan upplevas som en påtryckning eller överord om studiens tänkbara värde. Empati har en uppenbar roll i den etiska analysen. Det innebär att forskaren har förmåga till emotionell inlevelse och fantasi, kan ta deltagarens perspektiv och reflektera kring vad det kan innebära att medverka i studien samt är lyhörd inför deltagarnas upplevelser och reaktioner (Davis, 1983; Statens medicinsk-etiska råd, 2008).

Informationen ska finnas i skrift, och vid till exempel en ren enkätstudie ges den ofta bara i skriftlig form. Vid mer omfattande forskningsmedverkan, som vid en interventionsstudie, är det skrivna mer att se som ett första steg eller komplement till den information som ges muntligen. Det ska alltid finnas möjlighet att ställa frågor och få dem besvarade vid det muntliga informationstillfället och via kontaktuppgifter. Informationen ska vara lättillgänglig och anpassas till personens ålder och förutsättningar i övrigt. Vid forskning med barn ska information riktas till såväl barnet som till vårdnadshavaren.

Samtycke

Autonomiprincipen stipulerar att det är deltagarna själva som ska bestämma över sin medverkan i forskning. Det innebär att de har rätt att självständigt bestämma om de ska delta samt hur länge och på vilka villkor. De ska inte utsättas för påtryckningar att samtycka, och det bör inte föreligga något beroendeförhållande mellan forskaren och de tilltänkta deltagarna. Deltagarna ska när som helst kunna avbryta sin medverkan utan att bli ifrågasatta, och utan att det får några negativa följder för dem.

Samtycket ska vara informerat, frivilligt och *uttryckligt*. Att det ska vara uttryckligt innebär att inga underförstådda eller ”passiva” samtycken godtas. Om man enbart informerat om forskningen och ombett dem som inte vill delta att höra av sig är detta att betrakta som forskning utan samtycke. Ett sådant förfarande kan under vissa förutsättningar vara försvarbart, exempelvis om studien inte innebär någon aktiv intervention från forskningspersonerna och heller

inte omfattar några känsliga personuppgifter. Det betraktas dock inte som ett samtycke, eftersom ett sådant måste vara uttryckligt.

Ersättning till deltagarna?

En vanlig fråga är om det är etiskt försvarbart att erbjuda studiedeltagarna något slags ersättning. Man skulle kunna uppfatta ersättningen som en form av påtryckning och ett sätt att ”köpa” deltagarnas samtycke och medverkan. Å andra sidan kan den ses som en symbolisk motprestation, ett sätt att ge deltagarna erkänsla för den tid och kraft som de ägnar projektet. Man bör i alla händelser vara restriktiv med att erbjuda ersättning och den bör inte ligga på en nivå som skulle locka annars ointresserade deltagare. Den klassiska biobiljetten är i den meningen oproblematisk; den kan uppskattas som en trevlig gest men innebär knappast en påtryckning.

Det är i allmänhet också oproblematiskt att ersätta studiedeltagare för resor och andra kostnader som deras medverkan har medfört. Det har även förekommit att föräldrar erbjudits hjälp med barntillsyn för att underlätta deras deltagande i en studie av föräldrastöd. Det är sannolikt att man därmed har lyckats rekrytera en del föräldrar som annars skulle ha tackat nej, men man kan nog utgå från att det inte är barntillsynen i sig som gjort dem intresserade av studien.

Vid etikprövning ska forskaren uppge om och i så fall i vilken form ersättning ska erbjudas undersökningsdeltagarna. Huvudregeln är att eventuella ersättningar utformas med hänsyn både till deltagarnas situation och vad studien kräver av dem. Det får inte uppfattas som att man köper deras medverkan.

Dokumentation av samtycket

Samtycket ska dokumenteras[5], och det vanligaste sättet att göra detta är att deltagaren skriver under en samtyckesblankett. Underskriften innebär att deltagaren intygar att hon eller han har fått infor-

5 Detta krävs numera enligt svensk lag, etikprövningslagen (SFS 2003:460), vilket är i överensstämmelse med internationella forskningsetiska koder och rättspraxis.

mation om forskningen och samtycker till att medverka under de föregivna premisserna – vilket inkluderar rätten att avbryta forskningsdeltagandet. Samtyckesblanketten är alltså inte ett bindande kontrakt för forskningspersonen, och det är forskarens ansvar att se till att det inte uppfattas så. Om man inte använder en samtyckesblankett ska samtycket dokumenteras på annat sätt, till exempel i en projektloggbok, av vilken det framgår när och till vem det muntliga samtycket gavs.

Vid rena enkätundersökningar är det vanligt att man anser att den som svarar på enkäten också samtycker till att medverka i forskning. En sådan enklare form av samtycke är särskilt gångbar om enkäten besvaras anonymt och inte kan kopplas ihop med andra personuppgifter. Är det däremot till exempel en hälsoenkät ställd till personer som tidigare medverkat i en interventionsstudie, för att se hur de mår på sikt, bör man vara mera omsorgsfull. Deltagarna måste få tydlig information om att deras svar kommer att kopplas ihop med tidigare uppgifter som forskaren har tillgång till, och de bör ges tillfälle att uttryckligen samtycka eller avböja till detta.

Ställföreträdande samtycke

Minderåriga barn företräds av sina vårdnadshavare då det formella samtycket ska inhämtas. Givetvis bör även barnet självt informeras om studien och lämna sitt medgivande på ett sätt som är anpassat till barnets ålder och mognad. Om barnet då motsätter sig forskningsmedverkan ska det respekteras. Enligt svensk lagstiftning får den som fyllt 15 år men inte 18 år, och inser vad forskningen innebär för hans eller hennes del, själv informeras om och samtycka till forskningen. Forskaren behöver då alltså inte inhämta samtycke från föräldrarna. Liknande praxis finns i andra länder. Här finns ett tolkningsutrymme: Vad menas med att ”inse vad forskningen innebär”? Vid mer omfattande studier, där man samlar in särskilt integritetskänsliga uppgifter, eller då forskningen kan förmodas påverka den unga eller dennes anhöriga, är det vanligt att man gör en restriktiv tolkning och även tillfrågar vårdnadshavarna.

Människor kan ha nedsatt beslutskompetens på grund av sjukdom, psykisk störning, åldersdemens eller annat försvagat hälsotillstånd. Det är exempel på tillstånd då det kan vara svårt att ta till sig forskningsinformation och förstå vad det skulle innebära att medverka. Dessutom är det då vanligt att man befinner sig i beroendeställning visavi en vårdgivare. I sådana lägen är det sällan möjligt att ta självständig ställning till sin forskningsmedverkan, det vill säga att samtycka. Samtidigt kan det finnas angelägen forskning som inte kan genomföras utan att sådana personer deltar, exempelvis studier som syftar till att utveckla nya behandlingar för en viss typ av ohälsa.

En studie där deltagarna har nedsatt beslutsförmåga etikprövas särskilt omsorgsfullt. Det ska inte gå att få motsvarande kunskap genom att studera fullt beslutskompetenta personer. Vidare ska forskningen kunna förväntas medföra något positivt för studiedeltagarna själva, eller för andra som lider av samma problem. Risken för skada ska bedömas som minimal. Forskaren bör samråda med de närmast anhöriga, och i förekommande fall också med god man eller annan förvaltare som är utsedd att företräda personens intressen. Så långt som möjligt ska deltagaren själv informeras om forskningen. Om denne själv, eller någon av dem som man samrått med, motsätter sig deltagande får undersökningen inte genomföras.

Samtycke i tvångssituationer

Forskning som bedrivs på människor som befinner sig i en tvångssituation är ytterligare känslig. De befinner sig i en beroendesituation, där det egna handlingsutrymmet redan är kraftigt begränsat. För att det ska vara etiskt försvarbart att tillfråga dem om att delta i forskning måste forskningen vara särskilt angelägen och förväntas medföra något positivt för studiedeltagarna själva eller människor i samma situation. Riskerna för skada måste te sig minimala. Information och samtycke måste hanteras med särskild omsorg så att autonomiprincipen beaktas i bästa möjliga mån.

Forskning i tvångssituationer är alltså etiskt problematisk, men

det betyder inte att den är ogenomförbar. I stället är det angeläget att man finner sätt att hantera den etiska problematiken. Det är en förutsättning, exempelvis för att utveckla bättre behandlingsmetoder för tonåringar som vårdas på särskilda ungdomshem.

Hur långt sträcker sig samtycket?

Samtycket baseras på den forskningsinformation som deltagaren fått ta del av – varken mer eller mindre. Förutom en beskrivning av vad själva interventionen innebär har deltagaren i en interventionsstudie också fått information om vilka för- och eftermätningar som ska göras. Kanske planerades redan från början en uppföljande mätning. Deltagaren är då förberedd på att bli kontaktad igen efter till exempel sex månader, alternativt införstådd med att forskaren efter en viss tid kommer att inhämta uppgifter via ett eller flera offentliga register. Uppföljningsstudier har utan tvekan ett stort vetenskapligt värde. I själva verket borde sådana göras i långt större utsträckning än vad som är fallet, men forskningsfinansieringen är sällan så långsiktig att det är möjligt.

Hur förhåller man sig då om man vill göra en registeruppföljning utan att uttryckligen ha informerat studiedeltagarna om detta? Huvudregeln är att de ska kontaktas igen, få information om den fortsatta uppföljningen och få tillfälle att samtycka eller avböja fortsatt medverkan. I vissa lägen kan det dock finnas risker med att ta förnyad kontakt. Låt oss säga att interventionen ligger långt tillbaka i tiden, och att deltagarna behandlades för en problematik som det kan vara smärtsamt att påminnas om. En uppföljning via Socialstyrelsens register skulle kunna ge en fingervisning om den behandlade gruppen har lägre ohälsotal över en tioårsperiod än en jämförbar grupp som inte fick behandling. Eftersom en registeruppföljning inte kräver deltagarnas aktiva medverkan, vore det alltså rent praktiskt möjligt att genomföra den utan deras samtycke. Men vad säger etiken?

Det är möjligt att i detta läge ansöka om att få göra en uppföljning *utan* samtycke. Ansökan kommer dock att bedömas långt mer

restriktivt än om det vore en ren registerstudie, i vilken forskargruppen aldrig har haft direktkontakt med forskningspersonerna. Det är just detta som är den springande punkten; det finns en tidigare explicit överenskommelse mellan studiedeltagare och forskare som inte inkluderar en registeruppföljning. Studiedeltagarnas samtycke och medverkan i ursprungsstudien har alltså byggt på vissa premisser som forskaren i det här läget vill sätta sig över. Om en uppföljningsstudie utan samtycke skulle få etiskt godkännande eller ej är avhängigt flera faktorer. De tyngst vägande är:

1. Om det verkligen finns goda skäl att avstå från att kontakta studiedeltagarna igen. Det kan vara fallet om deltagarna kan tänkas bli illa berörda av att påminnas om en tidigare svår period i livet, eller om det rör sig om en mycket stor kohort (sällsynt vid interventionsstudier).
2. Om registerinformationen har tillräcklig precision för att besvara centrala och angelägna forskningsfrågor. I många fall är registerinformationen ganska grovkalibrig, och det är tveksamt om uppföljning via register skulle ge ett egentligt nytt kunskapstillskott.
3. Om risken är liten att uppföljningen, i den utsträckning den blir känd, kan upplevas som ett otillbörligt integritetsintrång. Denna punkt är lika viktig som den är svårbedömd.

Kan jag avbryta forskningsdelen?

Individen har rätt att när som helst avbryta sin medverkan i forskning, utan att det medför några negativa konsekvenser. Enligt etikprövningslagen (19 §) har forskaren rätt att använda de data som dessförinnan har inhämtats från studiedeltagaren. Det innebär att forskaren kan inkludera dem i en ”intent-to-treat analysis” och för att analysera bortfall.

Men vad händer om en deltagare vill få fortsatt del av interventionen, men däremot inte vill besvara ytterligare formulär? Är det möjligt att hoppa av forskningen men kvarstå i interventionen? Nej, de kommer i ett paket, skulle nog merparten fors-

kare hävda – åtminstone de som bedriver modellutvärderingar.[6] Det är en begriplig ståndpunkt, om man strävar efter att presentera en välkontrollerad studie. Men ur etisk synvinkel är den inte oproblematisk. I synnerhet om det är kostsamt att få behandling på annat sätt, eller om det finns långa vårdköer, kan erbjudandet om behandling vid en forskningsinstitution vara mycket lockande. För den som sedan inte vill fortsätta delta i forskningen, kan det te sig som en otvetydigt negativ konsekvens att inte längre få tillgång till behandling. Det är viktigt att erkänna att studiedeltagare kan uppleva detta som en inskränkning i frivilligheten, och att de kan tycka sig stå i beroendeställning visavi forskaren.

Rättviseprincipen

Rättviseprincipen har två huvudbudskap. Den lyfter fram skyldigheten att inte diskriminera människor, vilket innebär att den ställer kravet att personer med lika behov ska behandlas lika. Det är alltså personens behov av en viss intervention som ska avgöra hur man handlar, inte personens kulturella bakgrund, kön eller sociala status. Rättviseprincipen innebär också en skyldighet att visa solidaritet med svaga eller särskilt sårbara individer, som på grund av sina egenskaper eller sin livssituation har ett ökat behov av omgivningens stöd.

Rättviseprincipen är i första hand formulerad för vård och behandling, och har en given plats då man diskuterar prioriteringar och resursanvändning. Vad innebär den då i forskningssammanhang? Rättviseprincipen manar till eftertanke kring vilken forskning som prioriteras, och vilka samhällsintressen och värderingar som styr valet av forskningsfrågor (Gambrill, 2011). På ett mer konkret plan uppmanar den forskaren till att reflektera över om de tilltänkta

6 Det här gäller i förstone modellutvärderingar där den experimentella kontrollen förutsätts vara hög. Vid verksamhetsutvärderingar har forskaren mindre möjlighet att upprätthålla en sådan.

studiedeltagarna behandlas på ett icke-diskriminerande sätt. Är till exempel informationen lika tillgänglig för alla? Solidaritet med dem som har ett ökat behov av stöd kan innebära särskild varsamhet i hur man hanterar frågor om information och samtycke. Det kan också motivera att man har en förhöjd beredskap att identifiera negativa reaktioner eller att erbjuda hjälp om behov av sådana uppdagas.

Konfidentialitet och datasäkerhet

Vid interventionsstudier behöver forskaren hantera stora mängder data. Enligt autonomiprincipen är det studiedeltagaren själv som bestämmer över sin medverkan, vilket betyder att de uppgifter som man lämnat till forskaren inte får användas på andra sätt än det man uttryckligen samtyckt till. Det innebär att alla personuppgifter ska förvaras på ett sådant sätt att ingen obehörig kan ta del av dem. Uppgifter om alla studiedeltagare ska ges största möjliga konfidentialitet.

Men kan man verkligen lova sina forskningspersoner detta? Nej, skulle professor Christopher Gillberg hävda. Han var huvudansvarig forskare för den långtidsuppföljning av barn med neuropsykiatriska diagnoser som stod i centrum för det omdebatterade ”Göteborgsfallet”. Den longitudinella studien innehöll mycket integritetskänsligt material, och forskargruppen hade utlovat de deltagande familjerna ”fullständig konfidentialitet”. Med hänvisning till detta löfte vägrade Gillberg och medarbetare att lämna ut materialet till två andra forskare, som hade för avsikt att göra en egen analys av materialet. Kammarrätten i Göteborg menade dock att de sistnämnda hade ”ett berättigat intresse” av att få del av det materialet, och att ett utlämnande, förenat med särskilt sekretessförbehåll, inte skulle medföra risk för skada eller men för forskningspersonerna.[7] Forskarna vägrade dock att följa Kammarrättens domslut och valde slutligen att förstöra hela det värdefulla forskningsmaterialet. Deras

7 Förbehåll enligt SekrL 14:9 avseende bland annat fortsatt sekretess. Brott häremot är förenat med straffansvar.

agerande ledde till straffansvar för tjänstefel för Gillberg (Warnling-Nerep, 2003; Rynning, 2009).

Göteborgsfallet visar att man inte kan utlova fullständig konfidentialitet, om man med detta menar att ingen utanför den egna forskargruppen kan ta del av data. Om man å andra sidan utgår från att deltagarna lämnat samtycke till att delta i forskning med ett visst syfte, kommer saken i ett lite annat läge. Data skulle även fortsättningsvis hanteras med sekretesskydd, i Göteborgsfallet liksom i andra fall då utomstående forskare granskar redan insamlade datamaterial. Kliniska forskningsmaterial kan alltså bli föremål för samma kritiska granskning som övrig forskning, vilket är en god sak för dess trovärdighet. I Göteborgsfallet kan man dock fråga sig om inte de båda utomstående forskarna hade ett syfte som låg alltför långt från det som deltagarna en gång samtyckt till.

Vid interventionsstudier samlas uppgifter om studiedeltagarna in vid upprepade tillfällen, vilket innebär att forskaren måste upprätta en databas för personuppgifter. Innehåller den information om deltagarnas hälsa – vilket snarare är regel än undantag vid interventionsstudier – handlar det definitionsmässigt om känsliga personuppgifter. Sådana måste hanteras enligt särskilda bestämmelser. Samtidigt som de måste skyddas från intrång, har varje studiedeltagare rätt att på begäran få ett utdrag med de egna personuppgifterna ur registret. För att leva upp till dessa dubbla krav måste man ha goda forskningsadministrativa rutiner och hög datasäkerhet.

Vanligen innehåller databasen inga uppgifter om namn och personnummer, eftersom dessa medger omedelbar identifiering. I stället förses varje forskningsperson med en personlig kod som utgör vederbörandes identitet i databasen, och kodnyckeln finns i säkert förvar hos projektledaren. Kodnyckeln anlitas då man för på nya data, och bevaras tills alla datainsamlingar är genomförda och alla data finns registrerade i databasen. Databasen kan alltså avidentifieras först då det inte längre är aktuellt att påföra några nya uppgifter. Har man förhoppningen om att få finansiering för en långtidsuppföljning måste man alltså bibehålla identitetskoder och kodnyckel.

Att själva databasen inte innehåller namn eller personnummer är alltså inte detsamma som att den är avidentifierad. Däremot är det tilltalande från integritetssynpunkt att data bearbetas ”anonymt”.

Jävsförhållanden och vetenskaplig redlighet

Det hör till ovanligheterna att interventionsstudier genomförs av helt oberoende forskare. Detta internationella fenomen är väl synligt även på hemmaplan. I allmänhet har någon i forskargruppen varit med och utvecklat interventionen, eller man har andra personliga investeringar i den. Detta är enbart ett konstaterande, och inte sagt som kritik mot dem som tar på sig den mödosamma uppgiften att pröva sin metod vetenskapligt. Säkerligen skulle många välkomna om en oberoende forskare ville genomföra en helt fristående studie. Det händer bara alltför sällan. Skälet till detta är troligen att interventionsstudier är så extremt tids- och resurskrävande. Det är svårt att få dem rimligt finansierade, och det tar lång tid innan man kan redovisa sina resultat. Det är då inget att förundras över att det främst är upphovspersonerna själva som är engagerade nog att sjösätta en studie.

Betyder det att interventionsstudier bedrivs av idel jäviga forskare, och därför ska betraktas med allmän skepsis? Nej, det vore naturligtvis en både felaktig och ofruktbar slutsats. Det faktum att forskaren inte är opartisk ställer dock extra höga krav på dennes vetenskapliga redlighet. Det innebär att den vetenskapliga metodiken ska vara adekvat och oförvitligt tillämpad, att data hanteras korrekt och redovisas ärligt och att materialet finns tillgängligt för granskning. Det innebär också ett särskilt ansvar i att inte tänja resultaten i positiv riktning och i förtid marknadsföra något som evidensbaserat.

Som motvikt till sådana frestelser kan man välja att öppet redovisa ekonomiska intressen, eventuell copyright och andra jävsförhållanden. Det utplånar visserligen inte forskarens personliga bias men erbjuder en viktig läsanvisning. I dag har det blivit allt vanligare att

vetenskapliga tidskrifter kräver att forskare ska deklarera sina bindningar, alternativt explicit uttrycka att inga sådana föreligger. En annan åtgärd för ökad transparens och trovärdighet är att publicera forskningsplanen i förväg.[8] Det är ett sätt att undvika att tillfälliga fynd retrospektivt lyfts upp till att utgöra svar på centrala frågeställningar.

Den amerikanska psykologen Carolyn Webster-Stratton, som utvecklat föräldraprogrammet *The Incredible Years (IY)*[9] (Webster-Stratton, 2001), har själv genomfört många effektstudier av programmet. Ursprungligen riktade sig IY till föräldrar vars barn behandlades inom barnpsykiatrin för utagerande beteendeproblem. Sedan dess har IY utvecklats vidare och använts som preventionsprogram på olika nivåer, med moduler som riktar sig till föräldrar, lärare och barnen själva. Programmet har fått stor spridning och forskare runt om i världen har prövat det i vetenskapliga studier. Webster-Strattons egna studier hör dock till de mest välkontrollerade och ett flertal av dem visar ingen eller åtminstone ingen entydigt positiv effekt av IY. Hon är exempel på att en upphovsperson mycket väl kan pröva en metod minst lika rigoröst som en oberoende forskare. Nyckeln är god vetenskaplig metodik, transparent redovisad. Även på hemmaplan har vi goda exempel på interventionsforskare som prövar egenutvecklade metoder i vetenskapliga studier av god kvalitet, tillgängliggör resultaten och avstår från att övertolka dem.

Frestelsen att övertolka sina resultat är inte förbehållen interventionsforskare, utan finns hos forskare inom alla områden. I själva forskarrollen ligger ett visst mått av tunnelseende och ett starkt engagemang i de valda frågeställningarna. Vetenskapliga texter är inte enbart objektivt redovisande, utan har också ett retoriskt anslag som bygger upp vissa argument och tolkningar (Gambrill, 2011). Mer eller mindre avsiktligt lyfts vissa påståenden fram på bekostnad av

8 För det samhällsvetenskapliga området finns www.ClinicalTrials.gov som är öppen för forskare från hela världen att registrera sina projektplaner.
9 På svenska benämnt *De otroliga åren*.

andra (Hilgartner, 2000). Det råder också en stenhård konkurrens om forskningsmedel som gör det extra lockande att överdriva den egna forskningens förtjänster.

Samtidigt bör man inte förneka att problemet är värt extra eftertanke inom interventionsområdet. Inte bara den som skapat en intervention, utan även de tilltänkta avnämarna och samhället i stort, kan ha ett kortsiktigt intresse av att överdriva enstaka positiva resultat. Interventionerna har ju syftet att komma till rätta med faktiska problem.

Ett näraliggande exempel är olika preventionsprogram som riktar sig till barn och unga. Trots otillräckligt vetenskapligt stöd för deras förebyggande effekter har ett flertal sådana program fått en mycket snabb spridning i svenska kommuner och landsting (SBU, 2010), och beslutsfattare runt om i landet framhåller att man erbjuder ”evidensbaserade metoder”. Det finns mycket som tyder på att den psykiska hälsan hos våra ungdomar gradvis försämrats under de senaste 20 åren, vilket utan tvekan är alarmerande. Om en begränsad intervention kan hjälpa till att vända trenden vore det minst sagt välkommet. Här, liksom inom hela interventionsområdet, finns det därför grogrund för en hel del önsketänkande. Interventionsforskare bär inte ensamma ansvaret för det – men genom vetenskaplig ärlighet kan de undvika att bidra till det. Det är bara så som kunskapen kan växa och som effektiva interventioner kan utvecklas och bli tillgängliga. Det om något är god forskaretik.

Forskningsetisk prövning

Etisk prövning av forskning som avser människor har sitt ursprung i den första Nürnbergkoden. Under andra hälften av 1900-talet kom etikprövning att bli en självklar del av medicinska forskares vardag. I Sverige inrättades forskningsetiska kommittéer vid landets medicinska fakulteter, och dessa kom med tiden att stå för ett slags forskningsetisk *peer review* av i princip all medicinsk forskning. Även större anslagsgivare inrättade etikkommittéer för att bevaka att de

projekt som beviljades medel levde upp till rådande etiska principer. Sedan flera decennier tillbaka kräver dessutom de flesta vetenskapliga tidskrifter att de studier som publiceras ska vara etikprövade och godkända.

Den svenska etikprövningslagen

Trots att etikprövning länge varit i det närmaste institutionaliserad, var det först år 2004 som den första svenska etikprövningslagen (EPL) infördes. Därmed var Sverige bland de sista inom EU att ge etikprövningen en författningsmässig grund. Europarådets konvention för skydd av mänskliga rättigheter och värdighet i tillämpningen av biologi och medicin, från 1997, har satt tydliga spår i den svenska lagen. Omkring millennieskiftet tillkom flera EU-direktiv, bland annat ett om läkemedelsprövning, som gjorde det allt mer nödvändigt att ge etikprövningen lagstöd.[10] EPL trädde i kraft 1 januari 2004.

Med EPL inrättades sex regionala etikprövningsnämnder samt en central nämnd som kan överpröva de regionala nämndernas beslut. Nämnderna består av både forskare och allmänrepresentanter och leds av en ”lagfaren domare”, det vill säga jurist. Lagen slår fast att forskaren ska få besked från nämnden inom 60 dagar efter det att ansökan lämnats in. Den svenska ordningen harmonierar med systemen i andra europeiska länder, vilket ju också varit avsikten.

Enligt EPL ska i princip all humanmedicinsk forskning etikprövas och godkännas av en regional etikprövningsnämnd, och prövningen ska ske innan projektstart. Även social- och beteendevetenskaplig forskning är prövningspliktig om den

- innebär behandling av känsliga personuppgifter (enligt 13 § personuppgiftslagen, PUL), eller personuppgifter om lagöverträdelser (enligt 21 § PUL), eller om den

10 I den 1998 införda personuppgiftslagen (PUL, SFS 1998:204) angavs att forskare kan få tillåtelse att behandla känsliga personuppgifter, under förutsättning att forskningsprojektet godkänts av en forskningsetisk kommitté. Dessa kommittéer var vid tiden för PUL:s införande och t.o.m. år 2003 helt oreglerade. Detta var ytterligare ett skäl till varför det var hög tid att införa en etikprövningslag.

- utförs enligt en metod som syftar till att påverka forskningspersonen fysiskt eller psykiskt, eller som innebär en uppenbar risk att skada forskningspersonen fysiskt eller psykiskt.

Interventionsstudier syftar till påverkan och är därmed alltid prövningspliktiga, enligt den andra punkten. Dessutom omfattar de mätningar som görs före och efter intervention i regel känsliga personuppgifter, och då ska studien även av detta skäl etikprövas.

Ett hinder som ska passeras?

Det är inte ovanligt att forskare sinsemellan talar om etikprövning som i första hand ett hinder som bör forceras på ett så smidigt sätt som möjligt. Det är inte så märkligt, då forskaren faktiskt är beroende av ett godkännande för att kunna bedriva sin forskning. Men det är bara ena sidan av saken. Ett annat, och mera fruktbart, sätt att se saken är att etikprövningen är en integrerad del av projektförberedelserna. Då man skriver en etikansökan tvingas man ta ställning till en rad praktiska frågor som rör projektet och datainsamlingen. Man får tillfälle att allsidigt tänka igenom vad mötet med de tilltänkta deltagarna kommer att innebära, från första kontakt igenom hela forskningsprocessen. Det är väl använd tid.

Om etikprövningsnämnden har några invändningar blir man vanligen anmodad att göra justeringar i proceduren. Det kan handla om tillägg i informationsbrevet, eller att en första kontakt bör tas på ett alternativt sätt. Ibland kan betänkligheterna vara mer omfattande, och då är det brukligt att forskaren får komplettera ansökan med olika slags förtydliganden innan nämnden fattar beslut. Det är ganska sällan som en ansökan avslås. Då det händer finns möjligheten till överklagan hos centrala nämnden. Den möjligheten finns för övrigt i varje ärende där forskaren är missnöjd med nämndens beslut. Vid överklagande är det inte ovanligt att centrala nämnden upphäver det tidigare beslutet. Ibland beror det på att forskaren gjort ytterligare justeringar i sin projektplan; andra gånger har den regionala etikprövningsnämnden helt enkelt gjort en för sträng bedömning. Om

etikprövningen slutgiltigt resulterar i avslag finns det goda och tydligt redovisade skäl till detta. Den forskare som inte passerade hindret har tack vare det förmodligen besparats en hel del bekymmer.

Checklista inför etikprövningen

Kapitlet har organiserats efter några centrala teman, där de fyra forskningsetiska huvudprinciperna utgör kärnan. Dessa teman är användbara tankeredskap då man reflekterar över forskningsetiska frågor. De sammanfattas nedan i form av en checklista, tänkt att användas då man förbereder en interventionsstudie och en etikansökan.

Principen att göra gott

- Är undersökningen välgrundad, det vill säga finns det goda skäl för hypotesen att interventionen kan ge positiv effekt i denna undersökningspopulation?
- Finns det skäl att göra pilotstudier?
- Är designen tillräckligt stark för att kunna besvara forskningsfrågan?
- Hur hanteras kontrollgruppen?
- Vilket ansvar har forskaren visavi intresserade som efter screening exkluderats från deltagande?

Principen att inte skada

- Tillhör undersökningsdeltagarna en sårbar grupp?
- Finns det ett rapporteringssystem för negativa utfall?
- Hur fångas negativa reaktioner upp?
- Finns en beredskap att erbjuda stöd eller vidare remiss om någon reagerar negativt?

Autonomiprincipen – information

- Hur ska informationen förmedlas? Skriftligt, muntligt?
- Om forskningen rör minderåriga eller andra personer med begränsad egen beslutskompetens, vilka ska informeras?
- Är informationen väl anpassad till respektive mottagare?

- Omfattar informationen alla de inslag som rimligen kan tänkas påverka villigheten att delta?
- Finns det tydlig information om kommande uppföljningar (särskilt viktigt vid registeruppföljning)?
- Är kontaktuppgifter till ansvariga forskare lättillgängliga?
- Är det tydligt att man när som helst kan avbryta sin medverkan utan att det får några negativa konsekvenser?
- Kan studiedeltagaren avböja fortsatt forskningsdeltagande men kvarstå i interventionen? Om inte bör detta framgå i informationen.

Autonomiprincipen – samtycke

- Finns det något i situationen som gör att studiedeltagaren kan uppleva en begränsning i rätten att bestämma över sig själv?
- Kan studiedeltagaren uppleva sig beroende av forskaren?
- Kommer studiedeltagaren erbjudas någon form av ersättning? Kan denna uppfattas som en påtryckning?
- Hur säkerställer man att samtycket verkligen är autonomt?
- Vem eller vilka ska samtycka? Är det aktuellt att vända sig till vårdnadshavare eller andra ställföreträdare?
- Hur ska samtycket dokumenteras?
- Är det tydligt vad samtycket omfattar – och inte omfattar?

Rättviseprincipen

- Kommer studiedeltagarna bemötas på ett icke-diskriminerande sätt?
- Är informationen lättillgänglig för alla?

Konfidentialitet och datasäkerhet

- Vilka typer av data kommer att samlas in, och hur kommer rådata att förvaras?
- Hur garanteras att ingen utomstående får tillgång till uppgifter om studiedeltagarna?
- Hur garanteras att personuppgiftslagen efterföljs?

- Hur ser datasäkerheten i övrigt ut?

Jävsförhållanden

- Finns det ekonomiska bindningar och intressen? Hur ska dessa redovisas?
- Finns det några hinder för att resultaten ska kunna kommuniceras fritt, i vetenskapliga tidsskrifter?

Slutord

I forskning som rör människor är forskningsetiska och metodologiska överväganden i grund och botten oskiljaktiga. Det gäller i synnerhet studier som berör människors vardag, och där forskaren har långvarig kontakt med deltagarna – vilket är utmärkande för interventionsforskning. Man gör det svårt för sig som forskare om man inte beaktar etiska frågor redan då man påbörjar planeringen av en ny studie. Idéer kring design, experimentell kontroll och mätmetoder behöver återkommande prövas mot etiska principer då forskningsplanen växer fram.

Fördjupningslitteratur

Beauchamp, T. & Childress, J. (2009). *Principles of biomedical ethics* (6th revised ed.). New York: Oxford University Press.

Hermerén, G. (1996). *Kunskapens pris: Forskningsetiska problem och principer i humaniora och samhällsvetenskap*. (Andra reviderade upplagan). Stockholm: HSFR.

Statens medicinsk-etiska råd (2008). *Etik – en introduktion. Etiska vägmärken 1* (omarbetad upplaga). Stockholm: Fritzes.

Referenser

Beauchamp, T. & Childress, J. (2009). *Principles of biomedical ethics* (6th revised ed.). New York: Oxford University Press.

Biederman, J., Faraone, S. V., Monuteaux, M. C. & Feighner, J. A. (2000). Patterns of alcohol and drug use in in adolescents can be predicted by parental

substance use disorders. *Pediatrics, 106,* 792–797.
Cavell, T. A. & Hughes, J. N. (2000). Secondary prevention as a context for assessing change processes in aggressive children. *J School Psychol, 38,* 199–235.
Davis, M. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and Social Psychology, 44*, 113–126.
Dishion, T. J. & Andrews, D. W. (1995). Preventing escalation in problem behaviors with high risk adolescents: Immediate and 1-year outcomes. *J Consult Clin Psycholog, 63,* 538–48.
Dishion, T. J., McCord, J. & Poulin, F. (1999). When interventions harm. *Am Psycholog, 54,* 755–64.
Dishion, T. J., Poulin, F. & Burraston, B. (2001). Peer group dynamics associated with iatrogenic effects in group interventions with high-risk adolescents. *New Dir Child Adolesc Dev, 91,* 79–72.
Gambrill, E. (2011). Ethical aspects of outcome studies concerning social, behavioral and educational interventions. Invited paper at the 2011 Stockholm Conference on Outcome Studies of Social, Behavioral and Educational Interventions.
Gillon, R. (2003). Ethics needs principles – four can encompass the rest – and respects for autonomy should be ”first among equals”. *J Med Ethics, 29*, 307–312.
Hayes, S. C., Luoma, J., Bond, F., Masuda, A. & Lillis, J. (2006). Acceptance and Commitment Therapy: Model, processes, and outcomes. *Behaviour Research and Therapy, 44*, 1–25.
Hermerén, G. (1996). *Kunskapens pris: Forskningsetiska problem och principer i humaniora och samhällsvetenskap*. (Andra reviderade upplagan.) Stockholm: HSFR.
Hilgartner, S. (2000). *Science on stage: Expert advice as public drama*. Palo Alto, CA: Stanford University Press.
Kagan, S. (1998). *Normative ethics*. Boulder, CO: Westview Press.
Palinkas, L. A, Atkins, C.J., Miller, C. & Ferreira, D. (1996). Social skills training for drug prevention in high-risk female adoelescents. *Prev Med, 25*, 692–701.
Petersson, B. (1994). *Forskning och etiska koder*. Nora: Nya Doxa.
Rynning, E. (2009). Privatlivet och forskningen – en dyster lägesbeskrivning. *Svensk Juridisk Tidskrift*, 566–584.
SBU. (2010). *Program för att förebygga psykisk ohälsa hos barn. En systematisk litteraturöversikt.* Stockholm: Statens beredning för medicinsk utvärdering (SBU). SBU-rapport nr 202.
Statens medicinsk-etiska råd (2008*). Etik – en introduktion.* (Omarbetad upplaga.) Stockholm: Fritzes.
Svensk tandläkaretidskrift (1948), s. 2.
Warnling-Nerep, V. (2003). Hemlig eller offentlig forskning om s.k. bokstavsbarn? Om talerätt i ärende om utlämnande av allmän handling, *Juridisk Tidskrift, 4*, 133–144.
Warren, K., Mober, D. P. & McDonald, L. (2006). FAST and the arms race: The

interaction of group aggression and the Families and Schools Together program in the aggressive and delinquent behaviors of inner-city elementary school student. *J Prim Prev, 27*, 27–45.
Webster-Stratton, C. (2001). The incredible years: Parents, teachers and children training series. *Residential Tretament for Children and Youth, 18*, 31–35.

Länkar

Etikprövning: www.epn.se
CODEX – Regler och riktlinjer för forskning: www.codex.uu.se
Svensk författningssamling: http://www.riksdagen.se/webbnav/index.aspx?nid=3910

Kapitel 4 | Lars-Göran Öst

Forskningsdesigner

För cirka tio år sedan kom det en student som var i slutet av sin psykologutbildning och sade att han ville bli doktorand hos mig. Jag frågade varför och hans svar var: ”Jag har hittat en så spännande design i en artikel.” ”Det var intressant”, sade jag, ”men vad har du för frågeställning?” Han såg ut som ett stort frågetecken och jag förklarade för honom att man måste utgå från en vetenskaplig frågeställning som är viktig att besvara och utifrån den väljer man forskningsdesign. Jag uppmanade honom att gå hem och tänka på det och återkomma när han hade en frågeställning, men han kom aldrig tillbaka.

Detta kapitel handlar inte om hur man hittar en intressant och viktig frågeställning utan om nästa steg i processen, vilka olika designer som framför allt är användbara i effektforskning inom det psykosociala området och frågor som hör ihop med dessa. Kapitlet börjar med en genomgång av fyra typer av validitet – intern validitet, extern validitet, begreppsvaliditet och statistisk beslutsvaliditet – och olika hot mot dessa. Därefter följer en beskrivning av designer för effektutvärderingar, med fokus på experimentella designer, olika typer av kontrollgrupper och alternativa jämförelsegrupper. Sedan beskrivs olika strategier för effektutvärderingar och mättillfällen i effektstudier. Kapitlet avslutas med en genomgång av poweranalys; hur man gör den och olika sätt att öka statistiska power i en

studie. Illustrationerna kommer både från min egen forskning och från internationella kollegor.

Olika typer av validitet

Vilka slutsatser som kan dras av en behandlingsstudie beror på hur väl forskaren har lyckats att tillgodose fyra olika typer av validitet.

Intern validitet handlar om i vilken utsträckning interventionen, snarare än ovidkommande faktorer, kan anses förklara de uppnådda resultaten. *Extern validitet* rör i vilken utsträckning resultaten kan generaliseras till andra människor, miljöer, mått etcetera än dem som gällde i den aktuella experimentella situationen. *Begreppsvaliditet* fokuserar på vilka specifika aspekter hos interventionen som var orsaken till resultatet, det vill säga vilken är den begreppsmässiga basen för effekten. *Statistisk beslutsvaliditet* refererar till i vilken utsträckning en relation mellan behandling och effekt demonstreras och hur bra studien kan upptäcka skillnader om de existerar i verkligheten.

För var och en av dessa typer av validitet finns det ett antal hot som kan utgöra trovärdiga alternativa tolkningar av studiens resultat. En kort genomgång av dessa följer här.

Hot mot den interna validiteten

Historia

Med historia avses varje händelse utöver behandlingen som inträffar som kan förklara resultaten. Det rör sig om effekter av händelser som är gemensamma för alla personer i deras vardagsliv (hemma, i skolan, på arbetet), till exempel terroristattacken den 11 september 2001. Inverkan av sådana händelser kan ändra klientens prestation och bli tagen för en effekt som är resultat av behandlingen, eller ha en negativ inverkan och minska effekten av behandlingen. Det är alltså viktigt att kunna skilja mellan effekten av händelser i personers liv och effekten av interventionen. Om man i intervjun efter behandlingen frågar klienterna om vilka olika livshändelser

som har inträffat är det möjligt att statistiskt analysera effekten av dessa.

Mognad

Med mognad menas processer inom personer som förändras över tid, till exempel att bli äldre, klokare, starkare, tröttare eller uttråkad. Mognad är ett problem endast om designen inte kan skilja ut effekterna av mognadsprocesser från behandlingen. Detta är särskilt viktigt i studier av barn och ungdomar, och en studie som inte har en obehandlad kontrollgrupp kan inte skilja ut behandlingseffekt från mognad.

Testning

Med testning avses de effekter som kan uppstå vid eftermätningen av att ha genomgått testet (olika mätningar) före behandlingen. I terapiforskning utförs mätning före och efter behandlingen för att kunna evaluera hur mycket en individ förändras över tid på ett specifikt mått. Förändringar vid eftermätningen kanske inte beror på behandlingen utan på effekter av upprepad mätning.

Instrumenteffekter

Med instrumenteffekter menas eventuella förändringar i mätinstrumentet eller själva mätproceduren över tid. Detta beror på att mätmetoderna, till exempel olika självskattningsskalor, inte är fullständigt reliabla (test-retest-reliabiliteten är mindre än 1.0) och att de personer som gör skattningar (t.ex. klienter, terapeuter, oberoende bedömare) kan förändra sina bedömningskriterier över tid. Då kan de förändringar som uppmäts bero på förändrade kriterier snarare än ändrat beteende hos klienterna. Här används begreppet *response shift* som innebär förändringar i en persons interna standarder för mätningen, till exempel en förändring i värderingar, perspektiv eller kriterier. En rekommendation är att forskaren bygger in en kvalitetskontroll av instrumenten och mätproceduren.

Statistisk regression

Begreppet statistisk regression innebär tendensen för extrema värden på ett mått att återgå mot medelvärdet av fördelningen när mätinstrumentet administreras en gång till. Eftersom man i behandlingsstudier ofta har ett minimikriterium vad gäller svårighetsgrad för att klienten ska inkluderas i studien innebär det att det är klienter med extremvärden som inkluderas. Om så är fallet kan man på statistiska grunder förutse att vid en andra testning kommer deras värden att ha gått mot medelvärdet i fördelningen, något som ofta sker i en obehandlad kontrollgrupp. För att minska effekten av detta hot är det viktigt att använda mätmetoder som har en bra test-retest-reliabilitet.

Selektionsbias

Med selektionsbias avses systematiska skillnader mellan grupper före behandlingen beroende på selektion eller tilldelning av klienterna till betingelserna. Detta uppstår i klinisk forskning när intakta grupper väljs till experiment- och kontrollgrupper på grund av att randomisering inte är praktiskt eller etiskt möjlig. Som exempel kan nämnas studier där man låter en avdelning på en psykiatrisk klinik vara experimentgrupp och en grannavdelning på samma klinik vara kontrollgrupp, eller en skolklass vara experimentgrupp och en grannklass vara kontrollgrupp. Om grupperna skilde sig åt redan före behandlingen kan de skillnader som uppmäts vid eftermätningen lika gärna bero på selektionsbias som en verklig behandlingseffekt.

Dessa sex hot mot den interna validiteten kan undvikas eller till största delen hanteras genom att klienterna randomiseras till experiment- (behandling) och kontrollbetingelse. Har man tillräckligt stort sampel så är det stor sannolikhet att dessa hot kommer att slå lika i de två betingelserna. Följande hot kan dock inte kontrolleras genom randomisering:

Bortfall

Med bortfall menas att vissa klienter som har inkluderats i effektutvärderingen avslutar sitt deltagande i förtid och före eftermätningen. Detta är ett hot mot den interna validiteten om det sker, men det utgör ett speciellt hot om det är differentiellt mellan grupper (t.ex. om olika betingelser är olika attraktiva) då det introducerar en selektionsbias även om randomisering gjorts. De kvarstående klienterna (s.k. *completers*) kan inte antas vara representativa för det ursprungliga samplet och grupperna kan inte antas vara ekvivalenta.

Det finns olika åtgärder en forskare kan vidta för att minimera detta hot. För det första bör behandlaren skapa en bra terapeutisk relation med sina klienter så att man tidigt kan snappa upp eventuellt missnöje och försöka anpassa behandlingen (inom givna ramar) till klientens önskemål. En designvariant är att inte ha ett fast antal terapisessioner som alla klienter ska genomgå utan ett intervall, till exempel 8–14, och att behandlingen kan avslutas när klienten är nöjd med den uppnådda effekten. På det sättet minskar risken för avhopp bland klienter som svarar snabbt på behandlingen och är symtomfria efter halva behandlingstiden. Vidare är det viktigt att försöka genomföra eftermätning även på de klienter som hoppar av behandlingen. I och med att alla klienter som randomiserats ska ingå i den statistiska analysen (s.k. *Intent-To-Treat; ITT*) är det bättre att få verkliga värden än att ersätta det saknade eftervärdet med olika former av imputering (jfr kapitel 11).

Kombination av selektion och andra hot

Hoten historia, mognad och instrumenteffekter kan interagera med betingelser så att de utgör ett nytt hot, till exempel selektion X historia. En av grupperna kan ha utsatts för en händelse som den andra gruppen inte har. För att utgöra ett hot mot validiteten måste skillnaden mellan betingelserna vara systematisk och skillnaden kan vara en trolig förklaring till resultaten. Bästa sättet att klara av detta hot är att inte göra studier där selektion utgör ett hot.

Diffusion eller imitation av behandling

Den behandling som ges till experimentgruppen kan oavsiktligt komma alla eller vissa klienter i kontrollgruppen till godo. Som exempel kan nämnas studier där till exempel KBT jämförs med rutinbehandling och samma terapeuter utför bägge behandlingarna. Terapeuterna kan anse att KBT är bättre och tycka synd om de klienter som inte får denna behandling. Att de då ”smyger in” KBT i rutinbehandlingen förekommer. Effekten av detta är att göra prestationen i experiment- och kontrollgruppen mera lika och alltså minska skillnaderna i studien.

Enklaste sättet att klara av detta hot är att alla terapisessioner spelas in på video eller dvd och att man under den handledning som terapeuterna har under studiens gång tittar på delar av sessionerna. Då kommer handledaren att upptäcka om det förekommer behandlingstekniker i en betingelse som inte ska vara där och ge korrigerande återkoppling.

Speciell behandling av eller reaktioner i kontrollgruppen

Det är inte ovanligt att personal tycker synd om klienter som randomiserats till kontrollgruppen och ger dem annan service, som mer övervakning av deras välmående eller speciella privilegier. Detta kan i sig vara en intervention som kan försvaga skillnaden mot effekten av den behandling som ges till experimentgruppen. Det kan också hända att klienter i kontrollgruppen blir demoraliserade eller börjar tävla när de vet att de utgör kontrollgrupp. Detta utgör ett hot om det slår olika i grupperna så att effekten av behandlingen försvagas.

Delvis kan detta hot hanteras genom att projektledaren noggrant dokumenterar att klienterna i kontrollgruppen inte får någonting som de inte ska ha. Problemet med demoralisering och tävlingsförsök kan inte kontrolleras utan får undersökas via den intervju som görs vid eftermätningen.

Hot mot den externa validiteten

Sampelkarakteristika

Sampelkarakteristika handlar om i vilken utsträckning resultaten kan generaliseras till andra klienter som varierar i ålder, etnisk bakgrund, utbildning etcetera jämfört med dem som ingick i studien. Ett problem i psykologisk grundforskning och framför allt i äldre behandlingsstudier är att psykologistudenter ofta är försökspersoner i studierna. Mycket talar för att dessa skiljer sig markant från till exempel vårdsökande klienter och att det därför blir svårt att generalisera resultaten av studier med så kallade analoga sampel. En variant av samma problem är att minoritetsgrupper är underrepresenterade i forskningsstudier. Detta har lett till att National Institute of Mental Health i USA kräver att det i anslagsansökan ska framgå hur forskaren planerar att rekrytera minoritetsgrupper till sin studie. En kritisk fråga är alltså om forskningsresultaten kan vara begränsade till att enbart gälla de grupper som ingår.

Ett problem som ofta diskuteras i samband med randomiserade kontrollerade studier (RCT) är att dessa använder sig av så många exklusionskriterier, till exempel att komorbiditet inte får förekomma, att man får fram ett väldigt homogent sampel av klienter som inte är representativt för de klienter som söker psykiatrisk öppenvård. Det som framför allt påpekas är den höga frekvensen av komorbiditet, det vill säga att klienterna uppfyller diagnoskriterierna för mer än en störning som förekommer i psykiatrin samtidigt, och att dessa klienter är svårare att behandla än homogena sampel utan komorbiditet. Huruvida komorbiditet är en negativ prediktor för behandlingsresultaten råder det inte konsensus om, men som forskare bör man trots det vara försiktig med att utesluta komorbiditet. Jag har i min terapiforskning inom ångestområdet genomgående endast exkluderat klienter om de vid sidan av sin ångeststörning har en svårare störning som det är viktigare att behandla, till exempel psykos, bipolär störning eller missbruk. Däremot har aldrig klienter med personlighetstörningar blivit exkluderade. Rent generellt kan

rekommenderas att samplet har en lämplig representation av klienter över olika karakteristiska som kan påverka utfallet.

Stimuluskarakteristika och miljöer

Detta hot handlar om i vilken utsträckning resultaten kan generaliseras till andra miljöer, intervjuare, terapeuter etcetera än dem som ingick i den aktuella studien. En viktig fråga är om resultaten från en RCT kan generaliseras till den kliniska rutinvården. En svårighet i forskningen är att identifiera vilka aspekter av olika miljöer som modererar eller interagerar med olika behandlingseffekter. En faktor som ofta diskuteras är att de terapeuter som arbetar i RCT skiljer sig från dem som arbetar i rutinvården. RCT-terapeuterna arbetar endast med en grupp av klienter och en eller två behandlingsmetoder och därför kan man anta att de blir duktigare på just detta än terapeuter i rutinvården. Ett annat hot som kommer in här är om en studie endast har en terapeut, vilket gör att man inte kan skilja terapieffekt och terapeuteffekt åt. Särskilt problematiskt blir det om studien jämför två behandlingsmetoder och har en terapeut som utför bägge, men där terapeuten har tydliga preferenser för den ena metoden, till exempel genom att man har utvecklat denna (Zettle, 2003).

Reaktivitet hos de experimentella arrangemangen och mätningar

Detta hot avser inverkan av klientens vetskap om att man ingår i en undersökning, vilket kan leda till en vilja att behaga terapeuten och att undvika svar som man tror kan innebära att intervjuaren bedömer en negativt.

Om klienterna är medvetna om att deras prestation mäts kallas måtten för påfallande (eng. *obtrusive*). Medvetenheten om att prestationen mäts kan ändra prestationen från det den annars är. Om detta leder till att klienterna beter sig annorlunda än de annars skulle göra kallas måttet reaktivt. Reaktivitet är dock en gradfråga och beror på vilken typ av mätmetod som vi pratar om. När det gäller

självskattningsskalor kan klienten naturligtvis helt och hållet styra hur man skattar och ge lägre skattningar av sina problem än man verkligen upplever för att behaga terapeuten. Beteendetest är svårare att påverka och psykofysiologiska mätningar eller hjärnavbildningsmetoder sannolikt omöjliga att påverka viljemässigt.

Interferens av multipla behandlingar

Detta problem handlar om att dra slutsatser om en viss behandling när den evalueras inom ramen för, eller samtidigt som, andra behandlingar. Den vanligaste situationen är att vissa klienter som ingår i en studie av psykologisk behandling samtidigt har en pågående psykofarmakabehandling. Ett alternativ i denna situation är att kräva medicinfrihet för att personen ska få delta i studien och under kontrollerade former sätta ut den aktuella medicineringen över 2–4 veckor. Detta krav leder dock till att vissa klienter, sannolikt de svåraste, inte vågar avsluta sin medicinering mot till exempel panikattacker. En vanlig reaktion är: ”Jag vet att jag har fyra panikattacker i veckan trots den medicinering jag har, men om jag slutar kanske jag får fyra attacker per dag och det klarar jag inte av.” För att kunna inkludera även dem i studien brukar man sätta upp två krav: (1) klienten ska ha stått på en konstant dos av läkemedlet under lika lång tid före behandlingen som den psykologiska behandlingen varar, till exempel tolv veckor, och (2) klienten lovar att inte ändra dosen eller byta till en annan medicin under behandlingsperioden. Om dessa krav uppfylls kan man värdera effekten av psykoterapin utöver en konstant medicinering.

När det gäller eventuell psykologisk behandling som klienten går i när han eller hon inkluderas i studien är det nödvändigt att klienten avslutar den, eller lägger den på is under den tid som man deltar i studien. Om detta inte görs kan det leda till negativ interferens och sämre resultat av den terapeutiska metod som studeras.

Nyhetseffekter

Detta hot berör möjligheten att effekterna av en behandling till viss

del beror på nyhetsaspekten. Massmedia slår ofta upp nya behandlingsmetoder med braskande rubriker, men utan att dessa ännu har genomgått någon seriös utvärdering. Ofta beskriver man en person som har lidit av sin störning i många år och prövat ett flertal behandlingar utan resultat – ända tills han eller hon träffade på den nya metoden X och blev helt återställd. Det är då naturligt att klienter med samma störning efterfrågar den nya och effektiva behandlingen X för sina problem.

En ny terapiform kan vara effektiv på grund av dess specifika procedurer eller dess nyhetsvärde. Nyhetseffekten är svår att värdera men det är lämpligt att forskaren försöker att dölja att en grupp får den nya behandlingen och en får den gamla vanliga (underförstått mindre effektiva). Sannolikt kommer dock nyhetseffekterna att minska över tid när den aktuella terapimetoden används av fler och fler terapeuter.

Testsensitisering

Detta hot innebär att administrering av en mätmetod före terapin kan sensitisera klienterna så att de svarar annorlunda på behandlingen. Klienter som mäts före behandlingen kan vara mer mottagliga för terapin än de som inte genomgår förmätning. Jämfört med klinisk rutinvård där mätningar före och efter behandlingen inte brukar utföras så kan studiernas förmätningar bidra till ett bättre behandlingsresultat. Även eftertestning kan sensitisera klienter till den behandling de har fått och ge resultat som inte skulle ha blivit fallet utan mätningen.

Tidpunkter för mätningarna

Resultatet av en terapistudie kan bero på vid vilken tidpunkt mätningen görs. En validitetsfråga är om samma resultat skulle uppnås om mätningen hade gjorts vid en annan tidpunkt, till exempel flera månader eller år senare. En terapi kan ge goda effekter vid mätning direkt efter avslutningen men effekterna står sig inte vid uppföljningen medan en annan terapi kan visa bättre effekter vid uppfölj-

ningen än vid eftermätningen. På grund av detta rekommenderas att forskaren alltid tar med uppföljningsdata, helst efter ett år, i sin artikel eller rapport. Detta gäller dock endast när två eller flera aktiva terapier jämförs med varandra. Av etiska skäl kan man inte låta en obehandlad kontrollgrupp fortsätta att vara obehandlad under en uppföljningsperiod på till exempel ett år.

Relationen mellan intern och extern validitet

Från prioriteringssynpunkt är ett experiments interna validitet viktigare än den externa validiteten. Man måste först visa ett otvetydigt resultat innan man kan fråga om dess generalitet. En väl utförd studie med hög grad av intern validitet visar vad som kan hända när experimentet arrangeras på ett visst sätt. Det är något helt annat att visa att interventionen har denna effekt och fungerar på detta sätt utanför experimentsituationen, i den kliniska rutinvården.

Hot mot begreppsvaliditeten

Uppmärksamhet och kontakt med klienter

Uppmärksamhet och kontakt som ges till klienter i experimentgruppen, eller olika uppmärksamhet i experiment- och kontrollgruppen, kan vara hot mot begreppsvaliditeten. Frågan är om det är de specifika komponenterna och procedurerna i den aktuella behandlingsmetoden som orsakar förändringen hos klienterna eller om det är gemensamma faktorer (eng. *common factors*) som leder till förändringen. Till de gemensamma faktorerna räknas till exempel den terapeutiska relationen, arbetsalliansen och förväntan om att behandlingen ska vara effektiv.

Stevens, Hynan och Allen (2000) presenterade en metaanalys av 80 originalstudier som hade undersökt de separata effekterna av gemensamma faktorer och specifika faktorer över olika utfallsmått och behandlingsmetoder. De fann att specifika faktorer gav signifikant större effektstorlek än gemensamma faktorer över olika mått.

Dessutom fann man att gemensamma faktorer inte gav en signifikant effekt för svårare störningar.

Placeboeffekten är verklig och kraftfull; man uppskattar att 30–40 procent av klienterna i olika studier förbättras av placebobetingelsen och att effekterna är ungefär lika stora av farmakologisk som psykologisk placebo (t.ex. Heimberg, Liebowitz, Hope, Schneier, Holt, Welkowitz m.fl, 1998). Medan man i farmakologiska studier kan ha trippel placebo, det vill säga varken klient, terapeut eller oberoende bedömare vet om den enskilda klienten har placebo eller aktiv drog, så kan man i psykologisk placebo endast ha dubbel placebo; både klient och oberoende bedömare kan vara blinda, däremot inte terapeuten.

Hofmann och Smits (2008) publicerade en metaanalys av placebokontrollerade studier av KBT vid ångestsyndrom och fann en total effektstorlek på $d = 0.73$. Störst placeboeffekt, det vill säga minst skillnad mellan aktiv terapi och placebo, fann man vid paniksyndrom (0.35) och generaliserat ångestsyndrom (0.51). Minst placeboeffekt, det vill säga störst skillnad till den aktiva behandlingens fördel fann man vid tvångssyndrom (1.37) och akut stressyndrom (1.31).

Enstaka operationer och smal stimulussampling

Exempel på detta hot är när två teoretiskt skilda terapier jämförs och terapeut ett ger terapi A och terapeut två ger B. Då är det en fullständig sammanblandning mellan terapeut och terapimetod och om det blir en skillnad i resultat kan man inte avgöra om den beror på att den ena terapimetoden är bättre än den andra. Det kan lika gärna bero på att den ena terapeuten är skickligare i sitt arbete än den andra. Om man jämför två teoretiskt skilda metoder bör man ha många terapeuter, minst fem per metod, så att skicklighetsfaktorer tenderar att jämnas ut.

En annan problematisk design kan vara att två teoretiskt liknande terapier, till exempel två former av KBT, jämförs och en terapeut utför bägge behandlingarna. Då hålls terapeutfaktorn konstant över

terapierna, men terapeuten kan vara mer trovärdig, bekväm, kompetent och effektiv med den ena terapin. Begreppsvaliditeten ökar med flera terapeuter. Samtidigt bör man se till att randomisera klienterna på terapeuter inom betingelsen och ha tillräckligt många klienter (minst tio) per terapeut för att kunna statistiskt analysera terapeuteffekten skild från terapieffekten (t.ex. Clark, Ehlers, Hackmann, McManus, Fennell, Grey m.fl., 2006).

Experimentledarens förväntningar

I både laboratorieforskning och klinisk forskning är det fullt möjligt att förväntningar och önskningar om resultaten hos experimentledaren påverkar hur klienter presterar. Förväntanseffekter är ett hot mot begreppsvaliditeten då de utgör en plausibel alternativ tolkning av effekterna som annars tillskrivs behandlingen.

I terapiforskning är det, förutom projektledare, både terapeut och oberoende bedömare som kan ha förväntningar på resultaten. Det är svårt att tänka sig en professionell terapeut som inte har en positiv förväntan på den behandling han eller hon bedriver utifrån tidigare forskning och kliniska erfarenheter. Det viktiga här är att terapeuten inte aktivt påverkar klienten att vid eftermätningen framställa sig i bättre dager än han eller hon är för att behandlingen ska framstå som så effektiv som möjligt.

När det gäller oberoende bedömare är det ytterst viktigt att forskningsledaren vidtar ett antal åtgärder för att försäkra sig om att bedömaren verkligen är ”blind” och oberoende. För det första bör man rekrytera bedömare som har rätt grundutbildning för att göra skattningarna, se till att de inte arbetar på den klinik eller mottagning där studien görs och dölja design och hypoteser för dessa. Sedan bör bedömarna tränas så att de uppfyller minimikraven på interbedömarreliabilitet och denna ska testas kontinuerligt under studiens gång. Slutligen bör man testa blindheten hos bedömarna genom att låta dem gissa vilken behandling (om någon) som den intervjuade klienten har fått. Om bedömare gissar rätt signifikant oftare än slumpen är han eller hon antagligen inte blind.

Signaler i den experimentella situationen

Implicita krav (eng. *demand characteristics*) inkluderar sådana källor till påverkan som information till klienterna innan de kommer till studien, specifika instruktioner och procedurer under behandlingen. När många komponenter som ges till behandlingsgruppen skiljer sig från dem som ges till kontrollgruppen kan implicita krav komma in som ett hot mot begreppsvaliditeten.

Hot mot den statistiska beslutsvaliditeten

Innan jag går in på hoten mot den statistiska beslutsvaliditeten är det några begrepp som behöver definieras. Dessa är:

- *Alfa* (α), sannolikheten att förkasta nollhypotesen, det vill säga ingen signifikant skillnad mellan experiment- och kontrollgruppen, när den är sann. Detta kallas också typ I-fel.
- *Beta* (β), sannolikheten att acceptera nollhypotesen när den är falsk. Detta kallas också typ II-fel.
- *Statistisk power*, sannolikheten att förkasta nollhypotesen när den är falsk, eller sannolikheten att finna en signifikant skillnad mellan betingelserna när dessa de facto skiljer sig åt. Denna sannolikhet är 1 - β.
- *Effektstorlek*, skillnaden mellan olika betingelser uttryckt som ett gemensamt mått över mätmetoder och över studier.

Låg statistisk power

Det vanligaste hotet mot statistisk validitet är låg statistisk power, eller låg sannolikhet att upptäcka en skillnad om det verkligen finns en i populationen. Låg statistisk power fördröjer teoretiska och empiriska framsteg och tar forskningsresurser som kunde användas på bättre sätt.

Statistisk power i en terapistudie är en funktion av:

- kriteriet för statistisk signifikans (α)
- samplets storlek (N)
- skillnaden mellan grupperna (effektstorlek).

Det man som forskare kan påverka är samplets storlek och effektstorleken. Under rubriken Poweranalys nedan beskrivs vissa åtgärder som kan användas.

Variabilitet i procedurerna

Ett annat hot mot den statistiska beslutsvaliditeten är variabilitet i procedurerna. Den består till exempel av individuella skillnader bland klienter beroende på hur inklusions- och exklusionskriterier administreras, slumpmässiga fluktuationer i deras prestationer på olika test och skillnader mellan experimentledare eller terapeuter i hur de administrerar intervjuer eller behandlingar. Dessa former av ovidkommande variation kan minimeras genom att vara noga med hur studien utförs, till exempel hur man väljer ut, tränar och handleder terapeuter, bedömare och forskningsassistenter, och hur mätningar görs och behandlingar utförs. Det är också viktigt att kontinuerligt mäta behandlingsmetodernas integritet, det vill säga terapeuternas följsamhet till manualen och deras skicklighet i utförandet av terapierna.

Heterogenitet hos klienter

Klienter kan variera i många dimensioner, till exempel kön, ålder, bakgrund, etnicitet med mera. Ju större heterogenitet det är i klientkarakteristika, desto mindre sannolikhet är det att upptäcka skillnader mellan betingelserna i studien. För att detta hot ska bli aktuellt måste heterogeniteten finnas på karakteristika som är korrelerade med effekterna av terapin. Det man som forskare kan göra är antingen att välja homogena sampel, vilket begränsar den externa validiteten, eller heterogena sampel som är tillräckligt stora så att effekten av klientkarakteristika kan värderas. Här är det viktigt att randomiseringen har skett på ett sådant sätt att dessa variabler fördelas över betingelserna på ett adekvat sätt. För att minska inverkan av detta hot är det viktigt att forskaren tydligt beskriver vilka inklusions- och exklusionskriterier som ska användas och verkligen följa dessa.

Bristande reliabilitet hos måtten

Om ett mått har bristande reliabilitet så kommer en större del av klientens poäng på instrumentet att orsakas av osystematisk och slumpvis variation. Detta leder till att relativt stor variabilitet kommer in i den statistiska evalueringen och effektstorleken tenderar att bli lägre.

Vid val av mätmetoder för en terapistudie är det viktigt att ta hänsyn till måttets reliabilitet; dels den interna konsistensen, dels test-retest-reliabiliteten. Den senare är särskilt viktig i behandlingsforskning där man utför upprepad mätning för att kunna undersöka förändring från före till efter behandling. Bristande reliabilitet blir således ett större problem vid upprepad mätning i och med att man då har två mätningar med reliabilitetsbrister.

Multipla jämförelser

Ju fler statistiska testningar som görs inom samma studie, desto större är risken att man finner en skillnad av slumpen (typ I-fel). Om man har ett α-värde på 0.05 så gäller det för en statistisk testning, inte för samtliga som görs. Gör man till exempel 100 testningar så kan man förvänta sig att fem är signifikanta av slumpen. När man gör multipla jämförelser är α betydligt större än 0.05, beroende på antalet test. Bonferroni-korrektion är ett konservativt sätt att lösa problemet, vilket dock förutsätter att testen är oberoende av varandra. Det man då gör är att dela 0.05 med antal test man utför. Om man till exempel gör fem test så blir α-värdet $0.05/5 = 0.01$ och om man gör 10 test $0.05/10 = 0.005$. Generellt bör man vara restriktiv med antalet statistiska testningar som utförs och ha tydliga hypoteser för utfallet som vägleder de statistiska analyserna.

Relationer mellan validiteter

Designfaktorer som gör ett experiment mer känsligt som ett test på relationen mellan oberoende och beroende variabel tenderar att begränsa generaliserbarheten hos fynden och omvänt egenskaper hos

ett experiment som ökar generaliserbarheten hos resultaten tenderar att öka variabiliteten och minska känsligheten hos det experimentella testet.

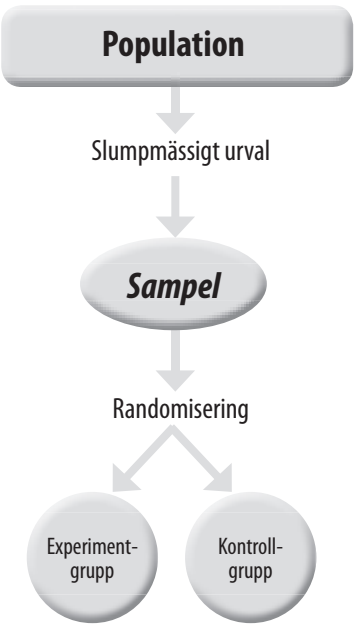
Designer för behandlingsforskning

Representativitet

När man gör en behandlingsstudie vill man kunna generalisera resultaten av denna utöver det sampel av klienter som har ingått i studien, vilket leder till frågan om samplets representativitet för populationen man vill uttala sig om. För att ett sampel ska kunna anses vara representativt för en population måste ett slumpmässigt urval ur denna dras på ett sådant sätt att varje individ i populationen har lika stor chans som varje annan individ att komma med i samplet. Detta förfarande kräver att varje individ i populationen är känd för forskaren.

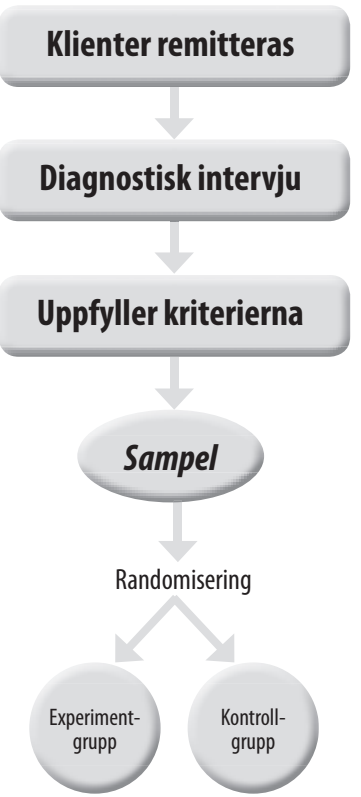
Anta att vi vill göra en behandlingsstudie av till exempel KBT vid egentlig depression hos vuxna och efter denna kunna uttala oss om hur KBT fungerar för klienter med denna diagnos. Generaliteten hos erhållna resultat beror på representativiteten hos försökspersonerna i relation till populationen som helhet.

För att varje individ i populationen vuxna med egentlig depression ska vara känd krävs det att varje vuxen person i befolkningen har genomgått en psykiatrisk diagnostisk intervju så att de som uppfyller diagnosen är kända och ingår i populationen som vi drar det slumpmässiga urvalet ifrån (figur 4:1). Så är inte fallet; det finns ingen psykiatrisk störning där förekomsten av störningen är känd i hela populationen. Inte ens om vi begränsar populationen till de personer som söker psykiatrisk öppenvård. Tvärtom visar Socialstyrelsens årliga rapporter om klienter som behandlats inom öppenvårdspsykiatrin att de i relativt stor utsträckning inte har fått någon diagnos när behandlingen avslutats, än mindre under pågående behandling (Socialstyrelsen, 2010).



Figur 4:1. Den teoretiska situationen då varje individ i populationen är känd och ett sampel kan skapas genom slumpmässigt urval ur population. Från samplet randomiseras klienterna till de olika betingelserna.

I realiteten går forskningen till på följande sätt: Forskaren informerar potentiella remissinstanser (t.ex. psykiatriska kliniker och öppenvårdsmottagningar) om att en behandlingsstudie om till exempel KBT vid egentlig depression kommer att startas vid en viss tidpunkt och pågå under en viss tidsperiod. De klienter som blir remitterade till projektet genomgår en noggrann diagnostisk intervju, till exempel med hjälp av SCID-I (First, Gibbon, Spitzer & Williams, 1996) för att fastställa om man uppfyller DSM-IV-kriterierna för diagnosen samt övriga inklusionskriterier som har ställts upp för studien. Därefter får klienten en beskrivning av studien och vad det innebär att delta i den, exempelvis randomisering till olika betingelser samt rättigheten att när som helst avbryta sitt deltagande. Om personen accepterar att delta får man skriva under ett informerat samtycke och därefter ingår man i samplet och randomiseras till en av de betingelser som ingår i designen (figur 4:2).



Figur 4:2. Verkligheten för psykoterapiforskning då samplet rekryteras genom att klienter remitteras till forskningsprojektet och därefter genomgår diagnostisk intervju. De klienter som uppfyller inklusionskriterierna utgör samplet och från detta randomiseras klienterna till de olika betingelserna.

Är samplet representativt för den teoretiska populationen?

Det sampel av deprimerade klienter som rekryteras på det sätt som visas i bilden via remisser är antagligen inte representativt för hela populationen om den vore möjlig att känna till. Om vi kunde låta samtliga personer i befolkningen diagnostiseras och därefter fylla i Beck Depression Inventory (Beck, Ward, Mendelson, Mock & Erbaugh, 1961) skulle vi antagligen få en fördelning som går från ganska låga till mycket höga poäng. Ett vanligt inklusionskriterium är att klienten ska ha minst 20 poäng på BDI, så att studien omfattar personer med tillräckligt svåra problem. Ett lika vanligt exklusionskriterium är att suicidala klienter inte ska ingå och riskera att ran-

domiseras till en kontrollgrupp som inte får behandling. Det innebär att samplet inte kommer att ha med de lättast och de svårast deprimerade som ingår i den teoretiska populationen, och därmed är samplet inte representativt.

I psykologisk forskning – klinisk, social eller annan – anses ett representativt sampel inte vara så nödvändigt och frånvaron av ett sådant anses inte vara ett stort problem. Man får i stället jämföra det aktuella samplet med vad man känner till om klienter med den aktuella diagnosen från tidigare forskning vad gäller bakgrundsdata, svårighetsgrad, antal tidigare depressiva episoder etcetera.

Experimentella designer

Det som kännetecknar experimentella designer är randomisering av klienter till betingelser. Detta innebär att klienterna i samplet tilldelas betingelser på ett sådant sätt att sannolikheten för varje individ att hamna i vilken grupp som helst är lika stor. Om vi har två grupper, en behandlings- och en kontrollgrupp, ska varje klient ha 50 procents chans att hamna i behandlings- respektive kontrollgruppen. Randomisering är nödvändig för att fördela karakteristika hos samplet osystematiskt över de olika grupperna. De faktorer som i tidigare forskning befunnits vara positiva eller negativa prediktorer för ett bra behandlingsresultat ska förekomma lika ofta i bägge grupperna. Om det före behandlingen är så att till exempel behandlingsgruppen gynnats med en övervikt av positiva prediktorer eller att kontrollgruppen missgynnats med en övervikt av negativa prediktorer, kan detta förklara skillnader mellan grupperna efter behandlingen.

Randomisering

Randomiseringen bör göras av en helt oberoende forskare som inte har med projektet att göra för att undvika eventuella misstankar om fusk i förfarandet. Det kan vara en person som arbetar vid ett annat universitet och som den aktuella projektledaren inte har något sam-

arbete med. Dessutom ska man göra hela randomiseringen innan någon klient har inkluderats i studien. Tidigare användes slumptalstabell men nu finns det enkla datorprogram på internet som med fördel kan användas, till exempel Research Randomizer (www.randomizer.org). I praktiken går det till på följande sätt:

- Antalet klienter som ska ingå i studien bestäms med hjälp av poweranalys, till exempel 60.
- Med datorprogrammet slumpas fram vilken grupp respektive löpnummer tillhör.
- Detta skrivs på numrerade lappar som läggs i numrerade kuvert (1–60), som försluts.
- Varje ny klient som inkluderas i studien får löpnumret som står på tur (1–60).
- När en ny klient har inkluderats, till exempel nr 27, e-postar projektledaren till den oberoende forskaren.
- Denne öppnar kuvert nr 27 och e-postar tillbaka om vilken grupp nr 27 ska tillhöra.

På detta sätt kan varken den person som intervjuar klienten för att bedöma om denna uppfyller inklusionskriterierna, klienten själv, terapeuten eller projektledaren påverka vilken grupp den enskilde klienten kommer att tillhöra. Denna modell kan också användas om man behöver använda en stratifierad randomisering för att försäkra att någon bakgrundsvariabel, till exempel kön, blir jämnt fördelad över betingelserna. Då gör man helt enkelt separata randomiseringar för män respektive kvinnor.

I följande beskrivning av olika experimentella designer förekommer vissa beteckningar som illustrerar designerna. Dessa betyder följande: R = randomisering, O = observation eller mätning och X = intervention eller behandling.

Pretest-posttest-kontrollgruppsdesign

Detta är den mest basala experimentella designen som innebär att

en grupp får behandling och en grupp inte får behandling. Det viktiga är att ingenting annat än X får skilja grupperna åt och att klienterna mäts före och efter behandlingen. Om det finns något annat än X som skiljer mellan grupperna kan detta utgöra ett hot mot den interna validiteten.

R	O_1	X	O_2
R	O_3		O_4

Fördelar med designen

Denna design har ett antal fördelar. Den kontrollerar för de vanliga hoten mot intern validitet. I och med att det finns mätning före och efter behandlingen så tillåter designen uttalanden om förändring efter behandlingen, något som terapeuter, klienter och vårdadministratörer är intresserade av att göra. Vidare tillåter designen (om tillräckligt många klienter ingår) utvärdering av effekten på olika nivåer av föreprestation, till exempel om behandlingen fungerar lika bra för dem som har måttlig respektive hög svårighetsgrad av problematiken. Vidare kan kraftfullare statistiska test användas om före- och efterdata tas med och bortfall kan analyseras bättre än utan förtestning.

Nackdel med designen

Den enda nackdel som finns hos denna design är en inverkan av förtestningen genom möjlig interaktion mellan testning och behandling, det vill säga att klienterna svarar bättre på behandlingen i och med att de har genomgått mätningar före än om de inte skulle ha gjort det.

Endast posttest-kontrollgruppsdesign

Denna design är i grunden densamma som den förra designen men förmätning saknas. Effekten av behandlingen bedöms endast på basis av eftermätningen. Denna design används i forskningssituationer där förtestning inte är önskvärd eller möjlig, till exempel om

man befarar att en mätmetod blir oanvändbar om den upprepas efter en viss tid.

R	X	O_1
R		O_2

Fördelar med designen

Denna design kontrollerar för de vanliga hoten mot intern validitet. Dessutom kan effekten av interventionen inte bero på initial testsensitisering, som är den enda nackdelen med pretest-posttest-kontrollgruppsdesignen.

Nackdelar med designen

Den primära nackdelen är att i klinisk forskning är det ofta kritiskt att känna till klienternas funktionsnivå före interventionen och gruppskillnader efter behandlingen kan vara skillnader mellan grupperna som fanns redan före interventionen. Dessa kan ha uppstått genom att vi hade otur med randomiseringen, vilket kan inträffa särskilt om det är ett litet sampel med kanske bara 8–10 klienter per grupp. Andra nackdelar är att man inte kan studera relationen mellan förevärden och förändring efter behandling, granska om eventuellt bortfall är differentiellt, det vill säga skiljer mellan grupperna, och man får en minskad statistisk power eftersom upprepad mätning inte är möjlig.

Faktoriella designer

I de två tidigare beskrivna designerna kan man undersöka en variabel i taget (behandling jämfört med ingen behandling eller en annan behandling). Det finns en annan typ av designer som tillåter samtidig undersökning av två eller flera variabler (faktorer) i ett enda experiment och de kallas faktoriella designer. Inom varje faktor administreras två eller flera betingelser och klienterna randomiseras till respektive betingelse inom respektive faktor, såvida inte en av faktorerna inte går att randomisera, som till exempel kön. Det vik-

tigaste skälet till att utföra faktoriella experiment är att den kombinerade effekten av två eller flera variabler, det vill säga deras interaktion, är intressant.

Den enklaste faktoriella designen kan beskrivas som 2 × 2, vilket innebär att det är två faktorer med två lägen (betingelser) i varje faktor och totalt fyra grupper. Om vi exempelvis vill undersöka effekten av terapeutisk utbildning för två olika KBT-metoder vid paniksyndrom kan en 2 × 2-design användas. I den första faktorn, utbildning, har vi tio terapeuter, fem med grundutbildning och fem med påbyggnadsutbildning i KBT. I den andra faktorn har vi behandlingsmetod; antingen kognitiv terapi (KT) eller tillämpad avslappning (TA). Klienterna randomiseras sedan på någon av de fyra betingelser som kombinationen av de två faktorerna ger och vi kan undersöka effekterna av: (1) utbildning (får terapeuter med högre utbildning bättre effekt än de med lägre utbildning, över de två metoderna?), (2) metod (ger KT bättre effekt än TA, över de två utbildningsnivåerna?) och (3) interaktionen mellan utbildning och metod (t.ex. får terapeuter med hög utbildning bättre effekt med KT än med TA, får terapeuter med lägre utbildning bättre effekt med TA än med KT etc.?).

En något mer komplicerad design är 2 × 3, som ger sex grupper och ännu mer komplicerad är 2 × 2 × 2, som ger åtta grupper. Man kanske vill undersöka könsfaktorn i tillägg till den tidigare beskrivna 2 × 2-designen, vilket innebär att man för var och en av de fyra betingelserna har lika många män som kvinnor bland klienterna.

Fördelar med designerna

Fördelarna med faktoriella designer är att de är ekonomiska; olika variabler kan studeras med färre klienter. Vidare kan de evaluera effekterna av separata variabler i ett enda experiment och de kan ge unik information om de kombinerade effekterna av oberoende variabler.

Nackdelar med designerna

Den primära nackdelen är att antalet grupper i undersökningen

multipliceras snabbt när nya faktorer eller nivåer av en faktor läggs till utöver den basala 2 × 2-designen. Dessutom är det svårt att tolka resultaten på ett begripligt sätt när multipla variabler interagerar med varandra.

Multipel behandlingsdesign – crossover-design

Det som karakteriserar dessa designer är att varje behandling som undersöks ges till varje klient i studien. Separata grupper används så att de olika behandlingarna kan balanseras över klienterna, det vill säga att behandlingarna ges i olika ordning. Den grundläggande typen av multipel behandlingsdesign kallas för crossover-design och användes tidigt inom farmakologisk forskning.

Liksom i de tidigare beskrivna designerna randomiseras klienterna från samplet till betingelser som får två olika behandlingsmetoder men i olika ordning. Den ena gruppen startar med metod A och den andra med metod B. Efter halva behandlingstiden görs en ny mätning och sedan byter klienterna till den andra metoden. Den kritiska faktorn med denna design är att grupperna får behandlingarna (*X*A, *X*B) i olika ordning.

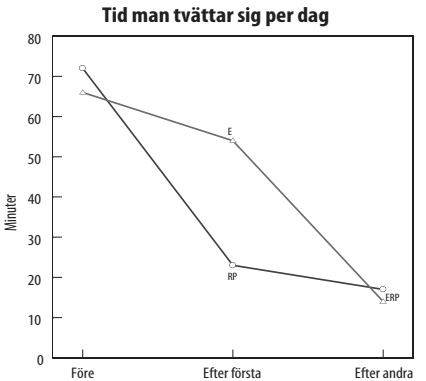
R	O_1	XA	O_2	XB	O_3
R	O_4	XB	O_5	XA	O_6

Fördelar med designen

Den primära fördelen med denna design är att man kan undersöka effekten av två olika behandlingar som ges i olika ordning till samma grupper av klienter.

Nackdelar med designen

När en crossover-design används inom psykofarmakastudier är det nödvändigt att efter den andra mätningen (O_2 och O_5 ovan) ha en så kallad wash-out-period på cirka två veckor då den första medicinen sätts ut. Om detta inte görs är det kombinationen av A och B respektive B och A som klienterna får i fas två.



Figur 4:3. Partiell crossover-design för att undersöka effekten av exponering (E) respektive responsprevention (RP) för klienter med tvättvång. Data hämtade från Foa m.fl. (1980).

Kan denna design verkligen användas i behandlingsforskning? Syftet med psykoterapi är ju att klienterna ska lära sig något viktigt som man har med sig i fortsättningen av livet, inte att man efter den andra mätningen ”glömmer bort” det man lärde sig i fas ett för att man ska kunna ta emot den andra behandlingen opåverkad av att ha fått den första. Det närmaste man kan komma denna design illustreras av en studie av Foa, Steketee och Mills (1980) där KBT-metoden exponering plus responsprevention (ERP) delades upp i sina två komponenter (se figur 4:3).

I fas ett fick den ena gruppen enbart exponering och den andra enbart responsprevention. Effekten var klart bättre för responsprevention än för exponering. När den andra mätningen utförts fick den första gruppen responsprevention i tillägg till exponering (som de fortsatte med) och den andra gruppen fick exponering i tillägg till responsprevention (som de fortsatte med). I fas två fick alltså bägge grupperna hela metoden ERP och det var ingen skillnad i effekt. Således kan man illustrera studien så här:

R	O_1	XA	O_2	XB+A	O_3
R	O_4	XB	O_5	XA+B	O_6

Kvasiexperimentella designer

Till skillnad från de experimentella designerna karakteriseras de kvasiexperimentella av att klienterna inte fördelas slumpmässigt på olika betingelser. Man startar vanligen med en grupp som har fått behandling och i efterhand (oftast) skapar man en kontrollgrupp genom att matcha klienter som så mycket som möjligt liknar dem som ingår i terapigruppen.

Pretest-posttest-design

Denna design liknar sin experimentella släkting med den avgörande skillnaden att klienterna i grupperna inte har slumpats fram.

Icke-R	O_1	X	O_2
Icke-R	O_3		O_4

Styrkan hos denna design beror helt och hållet på likheten mellan behandlings- och kontrollgruppen. Kontrollgruppen matchas med behandlingsgruppen genom att leta fram ”tvillingar” som i olika avseenden är lika de behandlade klienterna. Ett stort problem är att ju fler variabler man matchar på, till exempel kön, ålder, svårighetsgrad och problemduration, desto större är risken att man inte hittar tvillingar till alla klienter och vissa får exkluderas från samplet. Även om man lyckas få grupper som är lika på till exempel fyra variabler betyder detta inte att grupperna är ekvivalenta på alla dimensioner som är relevanta för behandlingen.

En variant av ovanstående design som ibland förekommer illustreras här:

Icke-R	O_1	X	O_2	
Icke-R			O_3	O_4

I denna design görs mätningarna av kontrollgruppen efter att behandlingsgruppen har genomgått eftermätning, vilket innebär att

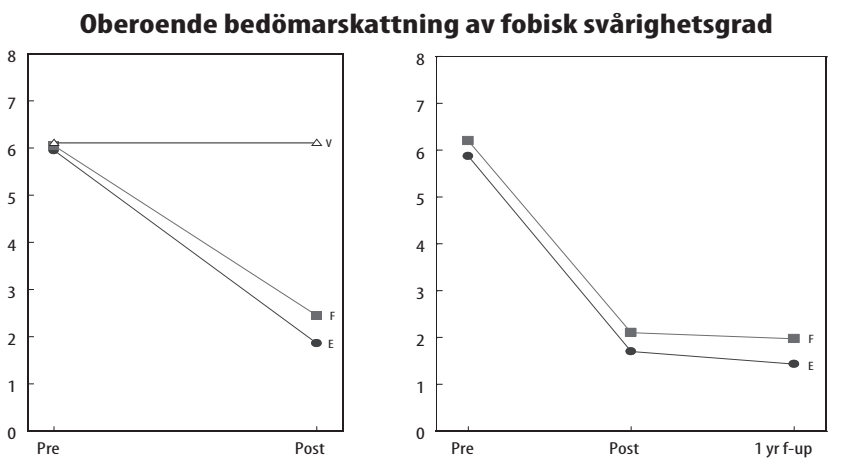
olika faktorer som funnits i miljön och kunnat påverka grupperna (positivt eller negativt) kan slå olika i de två grupperna.

Endast posttest-design

Den experimentella designen endast posttest-kontrollgruppsdesign har också en släkting bland de kvasiexperimentella designerna.

Icke-R	X	O_1
Icke-R		O_2

Det stora problemet med denna design, liksom den experimentella kusinen, är att gruppernas likhet före behandlingen inte kan bedömas i och med att ingen förmätning görs. För den kvasiexperimentella varianten gäller dessutom att grupperna kan vara mycket olika på flera variabler före behandlingen och att det blir mycket svårt att tillskriva eventuella skillnader mellan grupperna till interventionen.



Figur 4:4. Väntelistekontrollgrupp (VLK) för att undersöka två former av en-sessionsbehandling vid specifik fobi hos barn och ungdomar (E = barnet är ensamt med terapeuten, F = en förälder är med under terapin). Data hämtade från Öst, Svensson, Hellström och Lindwall (2001).

Olika typer av kontrollgrupper

Väntelistekontrollgrupp

Den vanligast förekommande kontrollgruppen i modern psykoterapiforskning är väntelistekontrollgruppen. Den innebär att klienterna randomiseras från samplet till en grupp som står på väntelista lika lång tid som behandlingen tar, vilket illustreras nedan, för att sedan få behandling. Denna kontrollbetingelse kontrollerar för spontanremission under förutsättning att klienterna vet om att de kommer att få behandling senare. Denna kontrollgrupp har dock vissa nackdelar, till exempel att klienterna som randomiserats till denna inte behöver anstränga sig på egen hand för att förbättras, bara vänta på att vänteperioden ska gå.

Ett exempel på denna typ av kontrollgrupp återfinns i figur 4:4.

I denna studie av Öst med flera (2001a) behandlades barn och ungdomar (7–17 år) som hade någon form av specifik fobi med så kallad ensessionsbehandling och randomiserades till tre grupper med 20 i varje. I en grupp var barnet ensamt (E) med terapeuten, i en annan grupp var en av barnets föräldrar med under behandlingen (F) och den tredje gruppen var en väntelistekontroll (VLK). Båda behandlingsgrupperna gav signifikant bättre effekt än ingen behandling, men det var ingen skillnad mellan dem (vänstra panelen av figur 4:4). När barnen i kontrollgruppen hade genomgått eftermätningen och ingen hade förbättrats signifikant randomiserades dessa på nytt; 10 till E- och 10 till F-gruppen, vilket totalt gav 30 i vardera betingelsen. Högra panelen visar att det inte är någon skillnad mellan grupperna och att resultaten står sig vid ettårsuppföljningen.

R	O_1	*X*	O_2		
R	O_3		O_4	*X*	O_5

Det förekommer ibland studier där väntelistegruppen inte ges behandlingsmetod X, utan en metod som den enskilde terapeuten själv

bestämmer, när klienterna ska behandlas i fas två. Detta leder till att man inte kan inkludera dessa klienter i uppföljningsmätningen och man har lägre statistisk power än man annars skulle ha.

Krav för att kombinera behandlings- och väntelistekontrollgrupp

För att man ska kunna kombinera den grupp som fick behandling direkt (T) och den som fick behandling efter att först ha stått på väntelista (VLK/T) anser jag att följande krav måste vara uppfyllda:

- T och VLK skiljer sig inte signifikant före terapin ($O_1 = O_3$)
- VLK-klienterna förbättras inte signifikant under tiden de är på väntelistan ($O_3 = O_4$)
- Förändringen efter terapin skiljer sig inte signifikant mellan T och VLK/T, d.v.s. ($O_1 \rightarrow O_2 = O_4 \rightarrow O_5$)

Dessa krav är uppfyllda i studien i figur 4:4.

Uppmärksamhets-placebo (ospecifik behandling)

Om den studie som forskaren designat syftar till att undersöka om det är den specifika behandlingen, eller unika karakteristika hos den, som är viktig för att ge förändringar bör en kontrollgrupp med uppmärksamhets-placebo användas. Klienterna randomiseras från samplet till en grupp som inte får veta om att de fungerar som kontrollgrupp och där behandlingen beskrivs som en vanlig terapimetod för störningen i fråga.

R	O_1	*X*	O_2
R	O_3	*UP*	O_4

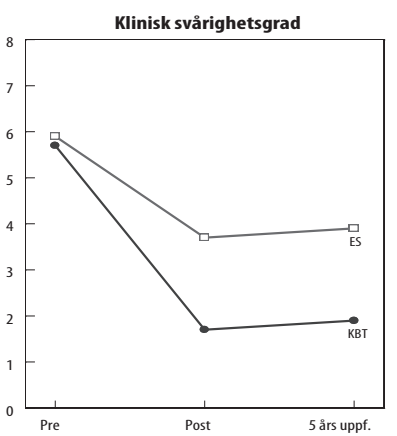
När forskaren bestämmer vad placebobehandlingen (UP) ska innehålla utgår man från den teori som anger vilka komponenter som är verksamma i behandling X. Dessa tar man bort från UP och det som återstår är att klienterna får samma kliniska omhändertagande som de i behandling X. De träffar en professionell terapeut som är

kunnig och skicklig, skapar en bra terapeutisk allians, är empatisk, varm, äkta etcetera. Ofta ingår också psykoedukation om störningen ifråga. Det är viktigt att samma terapeuter bedriver både X och UP, att man mäter klienternas tro på och förväntningar på den behandling de ska få och att terapitiden är densamma i bägge betingelserna. I figur 4:5 illustreras denna kontrollgrupp med en studie av KBT i grupp för social fobi.

Denna visar att placebogruppen, här kallad *Education Supportive therapy* (ES), fick en signifikant förbättring som dock var bara ungefär hälften av den förbättring som KBT-gruppen uppvisade. Andelen kliniskt signifikant förbättrade var 40 procent i placebogruppen mot 75 procent i KBT-gruppen.

Ingen-behandling-kontrollgrupp

En kontrollbetingelse som sällan används nuförtiden är ingen-behandling-kontrollgrupp. Den innebär att klienterna randomiseras från samplet till en grupp som inte får behandling, vare sig nu eller senare, vilket illustreras i rutan nedan. Denna kontrollgrupp kontrollerar för spontanremission, inklusive historia, mognad och regression. Skälet till att denna betingelse används allt mer sällan är



Figur 4:5. Design med psykologisk placebokontrollgrupp (education supportive therapy, ES) för att undersöka effekten av KBT vid social fobi. Data hämtade från Heimberg, Salzman, Holt och Blendell (1993).

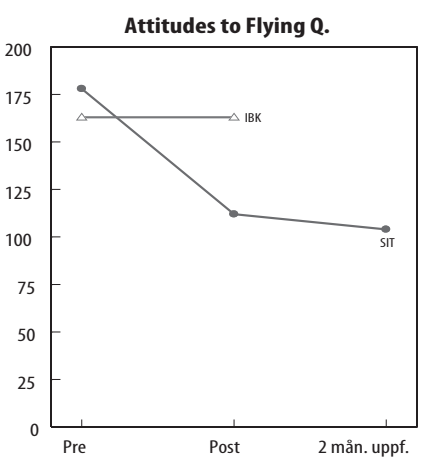
risken för bortfall i kontrollgruppen när man vet att det enda man får för att ställa upp på eftermätningen är ett tack, ingen behandling.

R	O_1	*X*	O_2
R	O_3		O_4

Denna kontrollgrupp användes i en studie av Beckham, Vrana, May och Gustafson (1990) rörande behandling av flygfobi där KBT-metoden *Stress Inoculation Training* (SIT) prövades. Av figur 4:6 framgår att kontrollgruppen inte förbättrades alls på självskattningsskalan för flygfobi medan SIT gav en signifikant förbättring som stod sig vid uppföljningen.

Ingen-kontakt-kontrollgrupp

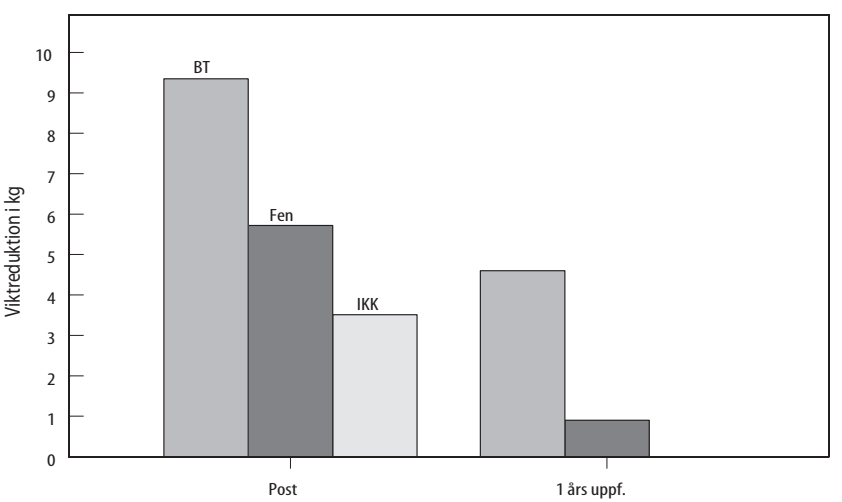
Om man har skäl att anta att klienterna påverkas av att man vet om att man ingår som kontrollpersoner i ett forskningsprojekt kan ingen-kontakt-kontrollgrupp användas. Det innebär att klienterna randomiseras från samplet till en grupp som inte får veta om att de fungerar som kontrollgrupp. Således måste förmätningen göras



Figur 4:6. Design med ingen-behandling-kontrollgrupp för att undersöka effekten av Stress Inoculation Training vid flygfobi. Data hämtade från Beckham m.fl. (1990).

dolt så att klienterna inte vet om att de utsätts för en mätning, vilket naturligtvis är svårt i psykologiska behandlingsstudier. Denna kontrollgrupp kontrollerar för ”självläkning” (spontanremission) och vetskap om att man deltar i ett forskningsprojekt.

Ett exempel på en studie med denna kontrollgrupp publicerades av Öst och Götestam (1976) rörande övervikt (figur 4:7). Beteendeterapi (BT) jämfördes med amfetaminderivatet Fenfluramin (Fen) och en kontrollgrupp (IKK) som inte visste om att de var kontrollgrupp. Detta kunde låta sig göras genom att samtliga personer som svarat på en tidningsannons om behandling av övervikt bjöds in till en föreläsning av en dietist om kostråd. Innan föreläsningen togs längd och vikt och vissa identifieringsdata på samtliga personer. De som hade ett BMI på minst 25 utgjorde ett sampel från vilket 45 personer (15 per grupp) randomiserades till de tre grupperna. Det var dock bara de två behandlingsgrupperna som informerades om vilken grupp de hamnat i medan kontrollgruppens deltagare fick tro att de inte hade kommit med i studien. När behandlingstiden var



Figur 4:7. Design med ingen-kontakt-kontrollgrupp för att undersöka effekten av beteendeterapi (BT) respektive Fenfluramin (Fen) vid övervikt. Data hämtade från Öst & Götestam (1976).

klar för de två aktiva grupperna kontaktades personerna i kontrollgruppen och informerades om att de kunde få beteendeterapi om de inom tre dagar kom in till försöksledaren för vägning. Därefter fick de fullständig information om upplägget och varför man inte hade talat om att de utgjorde en kontrollgrupp. Studien gjordes i början av 1970-talet och det är mycket osannolikt att en etisk kommitté skulle godkänna denna typ av design i dag. Resultatet visade att beteendeterapi gav signifikant större viktminskning än Fenfluramin, som i sin tur visade större viktminskning än kontrollgruppen. Vid ettårsuppföljningen hade personerna i bägge de aktiva behandlingarna ökat i vikt, men de som fått beteendeterapi var fortfarande bättre och den enda av grupperna som hade en signifikant lägre vikt än före behandlingen.

Alternativa jämförelsegrupper

Behöver man ha kontrollgrupper i alla studier? När det har gjorts ett antal kontrollerade studier av en viss behandlingsmetod och det är visat att metoden X ger signifikant bättre effekt än de olika kontrollbetingelserna behöver man inte fortsätta med kontrollgrupper. Då blir det i stället aktuellt med alternativa jämförelsegrupper.

Annan aktiv behandling

Denna design innebär att behandlingsmetoden av intresse jämförs med en annan aktiv behandling som tidigare visats ge god effekt för den aktuella störningen. Det är ingen idé att jämföra metoden X med en behandlingsmetod som visat sig vara effektiv för andra störningar men aldrig prövats för den aktuella störningen. Ta tvångssyndrom som exempel. Här är det ERP som har ett starkt evidensstöd och kognitiv terapi är möjligen effektiv. Däremot har aldrig tillämpad avslappning (TA), som är evidensbaserad för generaliserat ångestsyndrom och paniksyndrom, prövats för tvångssyndrom. Om den nya metoden X är utvecklad för tvångssyndrom så bör man jämföra med ERP och inte med TA.

Rutin- eller standardbehandling

En ofta förekommande jämförelsegrupp är rutin- eller standardbehandling. Det innebär att den nya behandlingen jämförs med standardbehandlingen (Treatment-as-Usual; TAU) i den aktuella kliniska miljön. Det är etiskt försvarbart i och med att alla klienter får behandling direkt; ingen behöver vänta på sin terapi. Från metodologisk synpunkt är det dock ett antal problem med denna jämförelsegrupp.

Problem

Även om en studie undersöker korttidsterapi (upp till 15–20 sessioner) tar det lång tid att rekrytera det antal klienter som ska ingå. En period på tre år för inklusion är inte ovanligt. För validiteten hos studien är det absolut nödvändigt att den sista klient som inkluderas år tre får samma behandling som den första fick år ett. Därför innebär TAU ett antal problem. För det första så *förändras* behandlingen ofta över tid vartefter terapeuterna går på workshops eller studiedagar och lär sig nya metoder, ofta hela eller delar av X. Det innebär att studien inte jämför X mot TAU utan X1 mot X2, vilket leder till lägre statistisk power och det blir därigenom svårare att visa signifikanta skillnader mellan betingelserna. För det andra så spelas sessionerna vanligtvis inte in, vilket innebär att terapeuternas följsamhet och kompetens inte kan mätas. För det tredje så får klienterna nästan alltid signifikant mindre terapitid än i den primära behandlingen (Öst, 2008). Dessa problem är vart och ett allvarliga hot mot en studies interna validitet och rekommendationen är att undvika rutinbehandling som jämförelsegrupp om man inte kan åtgärda dessa problem på ett tillfredsställande sätt. Dessutom finns det ingen kontroll för spontanförbättring.

Strategier för terapievaluering

När det gäller utvärdering av psykoterapier finns ett antal strategier (Kazdin, 2003). De vanligaste av dessa kommer att beskrivas och il-

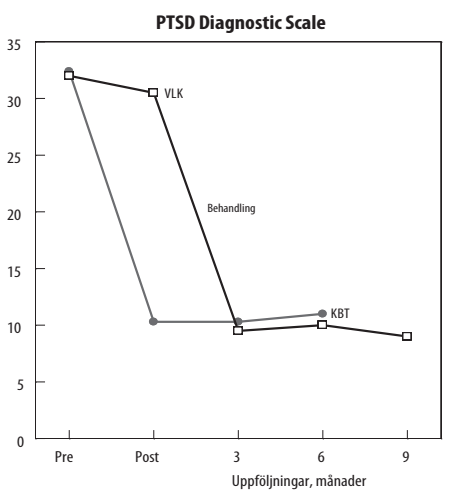
lustreras med exempel från min egen forskning. Det viktigaste att komma ihåg när det gäller designer för behandlingsforskning (och strategierna är specialvarianter av dessa) är att det alltid är den vetenskapliga frågeställning som forskaren har som styr vilken design eller strategi som är lämplig att använda.

Behandlingspaket

I början av utvecklingen av en behandlingsmetod vill man rimligtvis veta om den fungerar bättre än ingen behandling alls. Frågeställningen är således: Leder behandlingen till terapeutisk förändring?

För att besvara den frågan har man en design med två betingelser: behandling respektive ingen behandling eller väntelistekontrollgrupp. Strategin illustreras av figur 4:8.

Ehlers, Clark, Hackmann, McManus och Fennell (2005) jämförde KBT och väntelistekontroll (VLK) i en studie av klienter som hade utvecklat PTSD efter olika typer av trauman och fann en mycket god effekt av KBT medan VLK inte förändrades. När kontrollgruppen sedan fick KBT förbättrades den i samma utsträckning som den först behandlade gruppen och för bägge grupperna stod sig effekterna vid uppföljning efter sex månader. Slutsatsen av denna typ av



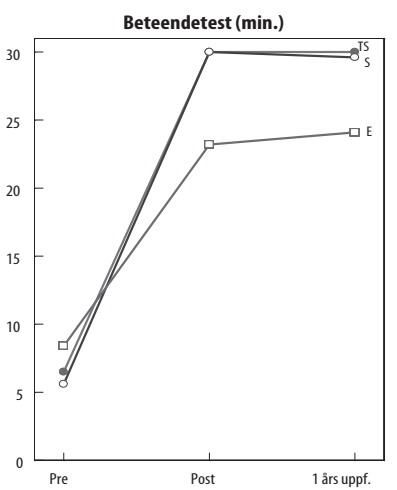
Figur 4:8. Behandlingspaketstrategi använd vid jämförelse av KBT och väntelistekontrollgrupp (VLK) för PTSD. Data hämtade från Ehlers m.fl. (2005).

studie är att behandlingen är effektiv men man vet inte vad det är i paketet som ger effekten.

Avrustning (eng. *dismantling*)

När man har visat att behandlingspaketet är effektivt (bättre än ingen behandling) kanske man vill undersöka vilka komponenter i paketet som är viktiga för att ge ett bra resultat. Den fråga som ställs är alltså: Vilka komponenter är nödvändiga, tillräckliga eller underlättande för att uppnå terapeutisk förändring?

För att besvara den frågan behöver man ha en design med två eller flera behandlingsgrupper som varierar i de behandlingskomponenter som ges. Strategin illustreras av figur 4:9. Öst, Fellenius och Sterner (1991) undersökte vilken av komponenterna i metoden tillämpad spänning för blodfobi som är viktigast för effekten; exponering eller spänningstekniken. I studien ingick tre betingelser: (1) tillämpad spänning, (2) exponering enbart och (3) spänning enbart, och alla fick lika lång terapitid, fem sessioner. Resultaten visade att enbart spänning och tillämpad spänning var lika bra (maxresultat på beteendetestet) och signifikant bättre än enbart exponering. I och med detta resultat blir slutsatsen att spänningstekniken är tillräcklig och exponering inte nödvändig för behandling av blodfobi.

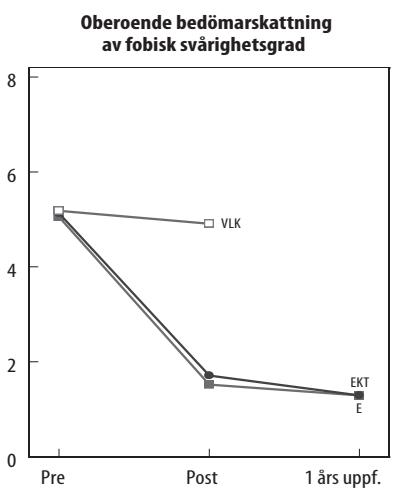


Figur 4:9. Dismantlingstrategi för att undersöka effekten av Tillämpad spänning (TS) och dess två komponenter Exponering (E) respektive Spänningstekniken (S) vid blodfobi. Data hämtade från Öst m.fl. (1991).

Konstruktiv strategi

Om man i forskning har funnit att den aktuella behandlingen är effektiv men att den inte fungerar för alla klienter, så blir en naturlig frågeställning: Vilka komponenter kan läggas till för att öka den terapeutiska förändringen?

För att besvara den frågan jämförs två eller flera behandlingsgrupper som varierar i behandlingskomponenterna (figur 4:10). Öst med flera (2004) undersökte om effekten av exponering vid paniksyndrom med agorafobi kunde bli mer kraftfull av att kompletteras med kognitiv terapi, eftersom denna metod visat sig effektiv vid panikattacker. Behandlingstiden var densamma och samma terapeuter utförde terapin i bägge betingelserna. I kombinationsgruppen fick terapeuten utifrån den enskilde klientens behov bestämma hur mycket av tiden som ägnades åt exponering respektive kognitiv terapi medan inga kognitiva tekniker fick användas i exponeringsgruppen. Resultatet var att bägge grupperna förbättrades signifikant mer än väntelistekontroll och att effekterna kvarstod vid ett års uppföljning.

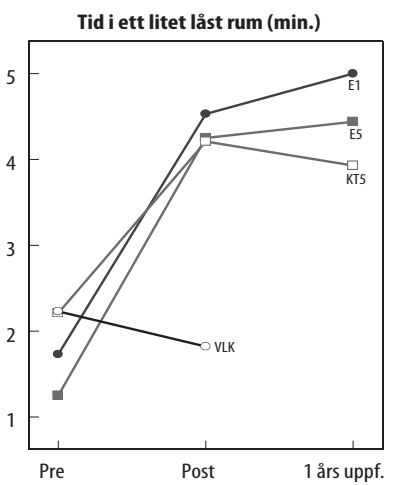


Figur 4:10. Konstruktiv strategi för att undersöka effekten av exponering plus kognitiv terapi (EKT) respektive enbart exponering (E) vid paniksyndrom med agorafobi. Data hämtade från Öst, Thulin & Ramnerö (2004).

Parametrisk strategi

Ett alternativ när behandlingen inte är tillräckligt effektiv är frågeställningen: Vilka förändringar kan göras i den specifika behandlingen för att öka dess effektivitet?

Frågan besvaras genom en design som har två eller flera behandlingsgrupper som skiljer sig i en eller flera aspekter av behandlingen. Oftast rör det sig om att man undersöker antalet sessioner och därigenom den totala behandlingstiden, men det kan vara andra aspekter, till exempel fördelningen av sessionerna i tid (t.ex. tolv sessioner en gång i veckan eller alla tolv under en vecka). Strategin illustreras av figur 4:11. I en studie av klaustrofobi jämförde Öst, Alm, Brandberg och Breitholtz (2001b) en session av exponering, fem sessioner av exponering, fem sessioner av kognitiv terapi utan någon exponering och VLK. Vid eftermätningen hade alla behandlingsgrupper förbättrats signifikant mer än VLK och det var ingen skillnad mellan dem. Vid ettårsuppföljningen var en-sessionsgruppen signifikant bättre än kognitiv terapi. Studiens resultat visar att man kan minska antalet sessioner från fem till en (och antalet timmar från fem till tre) utan att förlora i klinisk effektivitet vad gäller KBT vid klaustrofobi.

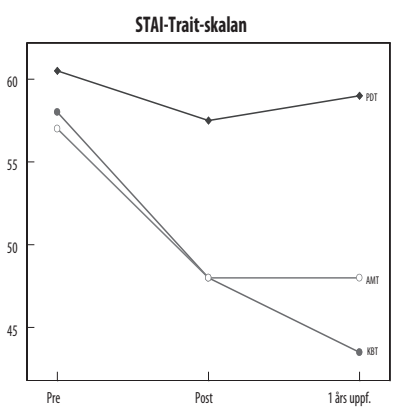


Figur 4:11. Parametrisk strategi för att undersöka effekten av exponering under en session (E1), under fem sessioner (E5), kognitiv terapi (KT5) respektive väntelistekontroll (VLK) vid klaustrofobi. Data hämtade från Öst m.fl. (2001b).

Jämförande effektstudie

Den frågeställning som de flesta förknippar med psykoterapiforskning är den så kallade hästkapplöpningen mellan KBT och psykodynamisk terapi (PDT) för en viss störning. Den frågeställning som man då har för sin studie är: Vilken behandling är mest effektiv för ett specifikt problem eller population?

För att besvara den frågan så jämförs två eller flera behandlingar (som ofta baseras på olika teoretiska grunder) för ett givet kliniskt problem (figur 4:12). Durham, Disher, Treliving, Hau, Richard och Stewart (1994) publicerade en studie av klienter med generaliserat ångestsyndrom (GAD) och jämförde psykodynamisk korttidsterapi (PDT), KBT och *Anxiety Management Training* (AMT). PDT och KBT utfördes av vardera två terapeuter med utbildning i och erfarenhet av sin respektive terapiform medan AMT utfördes av 13 psykiatrer efter en kortare utbildning i metoden. Resultatet visade att KBT och AMT gav signifikant bättre effekt än PDT och skillnaden ökade vid ett års uppföljning. Det är dock tveksamt att dra slutsatsen att KBT fungerar bättre än PDT för GAD-klienter i och med att det endast var två terapeuter inom respektive betingelse. Möjligheten att KBT-terapeuterna generellt var skickligare i att skapa en terapeutisk allians och i utförandet av terapin än PDT-terapeuterna kan inte uteslutas som en alternativ förklaring till resultaten, särskilt som inga mätningar gjordes av kompetens eller arbetsallians.



Figur 4:12. Jämförande effektstudie för att undersöka KBT, psykodynamisk korttidsterapi (PDT) och Anxiety Management Training (AMT) vid generaliserat ångestsyndrom. Data hämtade från Durham, Disher, Treliving, Hau, Richard & Stewart (1999).

Ett problem som ofta uppstår i dessa studier och som illustreras av Durhams studie är att man har för få terapeuter för respektive behandlingsmetod, vilket leder till att sådana faktorer som terapeutisk skicklighet och personlighet lätt kan spela in och det blir inte en rättvis jämförelse. För att denna strategi över huvud taget ska vara användbar måste man ha många terapeuter (minst fem) per metod och varje terapeut måste ha tillräckligt många klienter som randomiserats till dem (minst tio) för att terapeuteffekten ska kunna undersökas skild från terapieffekten.

Mättillfällen

Pre-post-mätningar

I effektforskning görs mätning före och efter behandlingen. I undantagsfall görs också mätning efter halva behandlingstiden. Det ger information som är nödvändig för att kunna evaluera effekterna av behandlingarna i de olika strategierna som beskrevs ovan. Det ger dock inte den nödvändiga informationen för att förstå behandlingseffekter, när och hur i terapiförloppet dessa inträffar etcetera. De flesta terapistudier har en fast behandlingsstruktur med små möjligheter till flexibilitet. Dessutom görs det ofta väldigt omfattande mätningar före och efter behandlingen och det finns en rädsla att klienterna tröttnar på alla mätningar och kanske hoppar av behandlingen. Trots detta har det på senare tid argumenterats för behovet av kontinuerlig mätning i gruppdesigner, på liknande sätt som görs i experimentella individ-designer (Barlow, Knock & Hersen, 2009).

Kontinuerlig mätning

I gruppdesigner skulle det vara värdefullt att integrera ytterligare mätningar under terapins förlopp, till exempel i början av varje session. Kliniskt har kognitiva terapeuter, till exempel i behandling av depression, använt sig av denna modell sedan 1970-talet. Klienterna får besvara *Beck Depression Inventory* i väntrummet före varje session och ta med sig formuläret till terapeuten som snabbt kan räk-

na ut den aktuella depressionsnivån (poäng på skalan).

En forskare som har tagit detta ett steg ytterligare är Michael Lambert vid Brigham Young University i Utah. Han har utvecklat *Outcome Questionnaire-45* (Lambert, Burlingame, Umphress, Hansen, Vermeersch, Clouse m.fl., 1999) som mäter tre faktorer: subjektivt obehag, interpersonella relationer och sociala rollprestationer. När klienten kommer till väntrummet får man en liten handdator som innehåller formuläret, och det tar cirka fem minuter att besvara frågorna. När man är klar med alla frågor och trycker på Enter-knappen skickas data trådlöst till terapeutens dator i behandlingsrummet och man får fram en graf som visar klientens poäng den aktuella dagen och samtliga föregående sessioner. I grafen finns det också linjer som visar om klienten i jämförelse med utgångsvärdet har försämrats, är oförändrad, har förbättrats eller har kommit innanför normalgruppens variationsområde.

Evaluering av förändringsmekanismerna

En huvudorsak till att utföra mätningar under terapins förlopp är studera de processer som är involverade i förändring och kunna beskriva verkningsmekanismen för behandlingsmetoden. Alla terapier har en begreppsmässig åsikt om varför klienter blir bättre i behandling, men det finns sällan starka evidens som stöder den åsikten. Vanligtvis saknas de nödvändiga mätningarna eller så har de gjorts på ett metodologiskt tveksamt sätt. För att kunna visa en kausal relation mellan den förmodade orsaken, till exempel den terapeutiska alliansen, och effekten måste denna (förändringen i alliansen) inträffa före utfallet (den terapeutiska förändringen). I de studier som mäter den förmodade orsaken och utfallet samtidigt, till exempel före och efter terapin, kan man inte visa den kausala relationen i och med att tidslinjen inte är korrekt. Se vidare kapitel 15 om mediatoranalys.

Uppföljningsmätningar

Mätning omedelbart efter terapin kallas eftermätning (post-treatment assessment). Varje senare mätning, oberoende av om den görs efter veckor eller år, kallas för uppföljningsmätning. Uppföljningsmätning ställer en viktig fråga för terapiforskning: Står sig förändringarna över tid? Denna fråga måste besvaras separat från frågan om behandlingen har initial effekt. En terapimetod kan ha en god effekt men inte stå sig tillräckligt bra över tid (klienterna har försämrats signifikant vid uppföljningen). Detta innebär inte att man ska förkasta metoden då det sannolikt är olika faktorer som gav de ursprungliga effekterna och som vidmakthåller de uppnådda effekterna på sikt. Bristande vidmakthållande av resultaten leder till frågan hur ett optimalt vidmakthållandeprogram (Öst, 1989) ska se ut för att resultaten ska stå sig på längre sikt.

Bortfall

Huvudproblemet vid uppföljning är bortfall, det vill säga att alla klienter som genomfört behandlingen och som det finns efterdata på inte kommer till uppföljningen. Öst (2009) presenterade en översikt av långtidsuppföljningar (189 studier) av KBT vid ångestsyndrom hos vuxna och fann att 86 procent av klienterna hade följts upp. Det visade sig vidare att för 86 procent av studierna hade behandlingseffekten stått sig i genomsnitt drygt två år efter avslutad behandling. Totalt hade 11 procent av de klienter som var kliniskt förbättrade vid eftermätningen återfallit vid uppföljningen medan 22 procent av dem som inte uppnått klinisk förbättring vid eftermätningen hade gjort det vid uppföljningen.

Det har visats i många sammanhang att ju längre uppföljningsperiod, desto större är bortfallet, framför allt på grund av att människor flyttar till en annan ort än där studien görs. Bortfall är ett problem för att uppföljningsdata kanske inte representerar den sanna funktionsnivån om alla klienter hade kunnat mätas. I detta sammanhang är det viktigt att forskaren gör en bortfallsanalys för att undersöka om de som kom till uppföljningen hade bättre initial be-

handlingseffekt än bortfallsgruppen och således inte är helt representativa. Klienternas samarbetsvillighet vid uppföljning kan bero på deras intryck av terapin, relationen med terapeuten och upplevd nytta av behandlingen. Över huvud taget är det önskvärt att få så litet bortfall som möjligt eftersom bortfall förstör den slumpmässiga sammansättningen av grupperna; betingelserna är inte längre jämförbara.

Praktiska beslut och alternativ

Det är ett antal frågor som forskaren behöver besvara inför uppföljningsmätning, till exempel: Ska klienterna komma till kliniken för en intervju (testning) eller kan den göras per telefon eller internet? En annan fråga är: Ska samma batteri av mätmetoder användas som vid pre- och post-mätningen eller kan ett mindre batteri av metoder räcka? Det är önskvärt att ha samma batteri som före och efter men följsamhet med uppföljning kan ökas markant genom att göra mätningarna så användarvänliga och korta som möjligt.

Poweranalys

Statistisk power handlar om i vilken utsträckning som en studie kan upptäcka en skillnad när den existerar. Alltför ofta ser man studier som har för låg statistisk power, kanske bara 25–30 procent, vilket kan anses som en form av ”gambling” med tanke på de stora kostnader som en modern terapistudie drar. Dessutom är det oetiskt att utsätta klienter för en studie där det inte är rimligt att komma fram till ett signifikant resultat då man redan från början har för låg power. Vilken nivå på power som är adekvat kan man inte komma fram till matematiskt. Beslutet baseras i stället på basis av en konvention om vilken skyddsmarginal man bör ha mot att acceptera nollhypotesen när den faktiskt är falsk (*beta*). Om power är .80 betyder det att forskaren har 80 procent chans att upptäcka en skillnad i sin studie om det finns en verklig skillnad i populationen.

Antalet klienter i studien

För att bestämma hur många klienter vi behöver inkludera i studien är det tre parametrar som måste bestämmas: alfa (p-värdet), power och effektstorlek (ES).

För att bestämma ES kan man använda sig av publicerade meta-analyser inom området, tidigare publicerade studier som har jämfört betingelserna av intresse på samma mått eller data från egen pilot-studie. Om man använder det sista alternativet och i sin studie ska jämföra behandlingsmetod X med en kontrollgrupp som inte behandlas, till exempel väntelistegrupp, gör man följande antagande: Kontrollgruppen kommer efter väntеperioden inte att ha förändrats signifikant utan ha ungefär det värde som behandlingsgruppen i pilotstudien hade som förevärde. Genom att jämföra med den gruppens eftervärde kan man estimera ES som

$$(M_b - M_k)/SD_{poolad}$$

där M_b och M_k är behandlings- respektive kontrollgruppens medelvärde vid eftermätningen och SD_{poolad} är den gemensamma standardavvikelsen för de två grupperna. Denna beräknas på följande sätt:

$$\sqrt{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2)/(n_1 + n_2 - 2)}$$

Om inget av ovanstående alternativ går att använda får man estimera ES på a priori-grund (underbyggd gissning). Här vet man från många metaanalyser av psykoterapiforskning att jämförelse mellan aktiv behandling och en obehandlad kontrollgrupp ofta ger ES = 0,80 och jämförelse mellan två aktiva behandlingar ofta ger ES = 0,20–0,40.

När man har fått fram sin ES kan man gå in i olika tabeller, till exempel s. 444 i Kazdin (2003), för att få fram hur många klienter som behövs per betingelse för att man ska ha en viss power (förslagsvis 80 procent) i sin studie. Det finns också kostnadsfria program på internet som kan användas, till exempel *G*Power 3.1.2* (www.psycho.uni-duesseldorf.de/aap/projects/gpower/) för att på ett mer detalje-

rat sätt räkna ut antalet klienter beroende på vilket statistiskt test som kommer att användas för dataanalyserna.

Variabilitet i data

För en given skillnad mellan grupperna på måttet så kommer ES att öka eller minska beroende på den standardavvikelse (SD) som skillnaden divideras med. SD kan vara större som en funktion av heterogenitet hos klienterna (till exempel vad gäller ålder, kön, socioekonomisk status etc.). Denna variabilitet är direkt relaterad till ES och statistisk signifikans och procedurer som reducerar ovidkommande variabilitet ökar den erhållna ES och powern.

Olika sätt att öka powern

Öka sampelstorleken

Det man först tänker på är att öka antalet klienter i studien. Detta är dock oftast svårt eller omöjligt i kliniska situationer. Man kan behöva rekrytera klienter från ett stort geografiskt område (multicenterstudie, som har sina egna metodproblem) eller fortsätta studien under väldigt lång tid, med risken att terapeuter och oberoende bedömare slutar sin anställning i projektet.

Här kan påpekas att om man endast ökar antalet personer i en betingelse (d.v.s. har olika stora betingelser) så får man endast en marginell ökning av power.

Öka förväntade skillnader mellan betingelserna

Viktiga frågor forskaren bör ställa sig i planeringen av sin studie kan vara: Är detta det starkaste testet mellan betingelserna eller kan hypotesen testas genom att göra en starkare manipulation eller en skarpare kontrast mellan betingelserna? I stället för att jämföra lite mot mycket (t.ex. rutinbehandling mot en ny aktiv behandling X) så är det en starkare kontrast med inget mot mycket (t.ex. väntelistekontroll mot X). Om man har ett givet antal klienter, till exempel 100, så är det bättre att fördela dessa på 2 betingelser om 50 i stället för 3 betingelser om 33, eller 4 betingelser om 25 klienter.

Använd förmätning

Fördelen med förmätning är att i olika statistiska test kommer feltermen för att evaluera ES att reduceras. Med upprepad mätning (pre- och posttest) kommer inomgruppsvariansen att tas med i beräkningen för att reducera nämnaren.

Den vanliga formeln för $ES = (M_1 - M_2)/SD$
vid upprepad mätning är den $ES = (M_1 - M_2)/SD\sqrt{1 - r^2}$
där *r* är korrelationen mellan pre- och post-mätning.

Om korrelation mellan för- och eftermätning är 0,10 så minskar nämnaren med en procent. Om den är 0,20 med två procent, 0,30 med fem procent, 0,40 med åtta procent, 0,50 med 13 procent, 0,60 med 20 procent, 0,70 med 29 procent, 0,80 med 40 procent och 0,90 med 56 procent. Detta illustrerar tydligt fördelen med att använda statistiska testningar som tar in förmätningen i analysen.

Variera alfanivåerna inom en studie

Ett alternativ är att höja α-nivån till exempelvis 0,10. Detta måste man bestämma i förväg (a priori) om man kan argumentera bra för det, till exempel om klassifikationen av grupper inte är perfekt, om måtten inte är väletablerade (tveksamma psykometriska egenskaper) eller om små effekter (skillnader) prediceras från början.

Använd riktade test

I mycket av klinisk forskning har man en uttalad åsikt om riktningen på förväntade skillnader. Forskaren vill troligen inte uttala sig om ES skiljer sig från noll utan om behandlingen är bättre än kontrollbetingelsen eller om behandling A är bättre än behandling B. Hypotesen som ska förkastas är inte tvåsidig (bättre eller sämre) utan ensidig (bättre). Högsta prioritet i behandlingsforskning är att utföra bra test på tydliga prediktioner. Ett riktat test på en riktad prediktion är troligen den bästa matchningen.

Minska variabiliteten i studien

Noggrann kontroll av variabiliteten kan uppnås genom att konstanthålla många variationskällor, till exempel att använda manualer för behandlingarna, tydliga instruktioner för hur olika mätningar ska göras, träning av oberoende bedömare, kontinuerlig mätning av deras interbedömaröverensstämmelse etcetera. Detta är exempel på åtgärder som leder till mindre variabilitet och därigenom starkare effektstorlekar.

Sammanfattning

I kapitlet har först en genomgång gjorts av olika typer av validitet som är viktiga vid effektforskning, vilka olika hot som är aktuella och förslag på hur man kan kontrollera eller klara av dessa. Därefter beskrevs olika typer av framför allt experimentella designer för behandlingsforskning, samt fördelar och nackdelar med dessa. Olika typer av kontrollgrupper och alternativa jämförelsegrupper beskrevs och illustrerades liksom specifika strategier som ofta används vid terapievaluering. Kapitlet avslutas med en genomgång av mättillfällen samt den mycket viktiga frågan om poweranalys; när och hur den ska göras och olika sätt som en forskare kan använda för att öka powern i sin studie.

Fördjupningslitteratur

Cook, T. T. & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Kazdin, A. E. (2003). *Research design in clinical psychology*, 4th ed. Boston: Allyn & Bacon.

Kendall, P. C., Butcher, J. N. & Holmbeck, G. N. (Red.) (1999). *Handbook of research methods in clinical psychology*. New York: Wiley.

Nezu, A. M. & Nezu, C. M. (Red.) (2008). *Evidence-based outcome research*. Oxford: Oxford University Press.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Referenser

Barlow, D. H., Knock, M. K. & Hersen, M. (2009). *Single case experimental designs*. 3rd ed. Boston: Allyn & Bacon.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J. E. & Erbaugh, J. K. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*, 561–571.

Beckham, J. C., Vrana, S. R., May, J. G., Gustafson, D. J., m.fl. (1990). Emotional processing and fear measurement synchrony as indicators of treatment outcome in fear of flying. *Journal of Behavior Therapy and Experimental Psychiatry, 21*, 153–162.

Clark, D. M., Ehlers, A., Hackmann, A., McManus, F., Fennell, M., Grey, N. m.fl. (2006). Cognitive therapy versus exposure and applied relaxation in social phobia: A randomized controlled trial. *Journal of Consulting and Clinical Psychology, 74,* 568–578.

Cook, T. T. & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Durham, R. C., Murphy, T., Allan, T., Richard, K., Treliving, L. R & Fenton, G. (1994). Cognitive therapy, analytic psychotherapy and anxiety management training for generalised anxiety disorder. *British Journal of Psychiatry, 165*, 315–323.

Durham, R. C., Disher, P. L., Treliving, L. R., Hau, C. M., Richard. K. & Stewart, J. B. (1999). One year follow-up of cognitive therapy, analytic psychotherapy and anxiety management training for generalized anxiety disorder: Symptom change, medication usage and attitudes to treatment. *Behavioural and Cognitive Psychotherapy, 27*, 19–35.

Ehlers, A., Clark, D. M., Hackmann, A., McManus, F. & Fennell, M. (2005). Cognitive therapy for post-traumatic stress disorder: Development and evaluation. *Behaviour Research and Therapy, 43,* 413–431.

First, M. B., Gibbon, M., Spitzer, R. L. & Williams, J. B. W. (1996). *User's guide for Structured Clinical Interview for DSM–IV axis I disorders*. Washington, D.C.: American Psychiatric Press.

Foa, E. F., Steketee, G. & Mills, J. B. (1980). Differential effects of exposure and response prevention in obsessive-compulsive washers. *Journal of Consulting and Clinical Psychology, 48*, 71–79.

Heimberg, R. G., Salzman, D. G., Holt, C. S. & Blendell, K. A. (1993). Cognitive-behavioral group treatment for social phobia: Effectiveness at five-year followup. *Cognitive Therapy and Research, 17*, 325–339.

Heimberg, R. G., Liebowitz, M. R., Hope, D. A., Schneier, F. R., Holt, C. S., Welkowitz, L. A. m.fl. (1998). Cognitive behavioral group therapy vs phenelzine therapy for social phobia: 12-week outcome. *Archives of General Psychiatry, 55,* 1133–1141.

Hofmann, S. G. & Smits, J. A. J. (2008). Cognitive-behavioral therapy for adult

anxiety disorders: A meta-analysis of randomized placebo-controlled trials. *Journal of Clinical Psychiatry, 69*, 621–632.
Kazdin, A. E. (2003). *Research design in clinical psychology*, 4th ed. Boston: Allyn & Bacon.
Lambert, M. J., Burlingame, G. M., Umphress, V., Hansen, N. B., Vermeersch, D. A., Clouse, G. C. m.fl. (1999). The reliability and validity of the Outcome Questionnaire. *Clinical Psychology & Psychotherapy, 3*, 249–258.
Socialstyrelsen (2010). Mot ett system för verksamhetsuppföljning på psykiatriområdet. Lägesrapport 2010. Stockholm: Socialstyrelsen, artikelnummer 2010-7-3.
Stevens, S. T., Hynan, M. T. & Allen, M. (2000). A meta-analysis of common factor and specific treatment effects across outcome domains of the phase model of psychotherapy. *Clinical Psychology: Science and Practice, 7*, 273–290.
Zettle, R. D. (2003). Acceptance and commitment therapy (ACT) vs. systematic desensitization in treatment of mathematics anxiety. *The Psychological Record, 53*, 197–215.
Öst, L-G. (1989). A maintenance program for behavioral treatment of anxiety disorders. *Behaviour Research and Therapy*, *27*, 123–130.
Öst, L-G. (2008). Efficacy of the third wave of behavioral therapies: A systematic review and meta-analysis. *Behaviour Research and Therapy, 46,* 296–321.
Öst, L-G. (2009). *Långtidsuppföljning av KBT vid ångeststörningar*. Föreläsning vid Beteendeterapeutiska Föreningens årsmöte i Uppsala, mars 2009.
Öst, L-G., Alm, T., Brandberg, M. & Breitholtz, E. (2001b). One vs. five sessions of exposure and five sessions of cognitive therapy in the treatment of claustrophobia. *Behaviour Research and Therapy*, *39,* 167–183.
Öst, L-G., Fellenius, J. & Sterner, U. (1991). Applied tension, exposure in-vivo, and tension-only in the treatment of blood phobia. *Behaviour Research and Therapy*, *29*, 561–574.
Öst, L-G. & Götestam, K.G. (1976). Behavioral and pharmacological treatments for obesity: An experimental comparison. *Addictive Behaviors*, *1*, 331–338.
Öst, L-G., Svensson, L., Hellström, K. & Lindwall, R. (2001a). One-session treatment of specific phobias in youth: A randomized clinical trial. *Journal of Consulting and Clinical Psychology*, *69*, 814–824.
Öst, L-G., Thulin, U. & Ramnerö, J. (2004). Cognitive behavior therapy vs. exposure in-vivo in the treatment of panic disorder with agoraphobia. *Behaviour Research and Therapy*, *42*, 1105–1127.

Kapitel 5 | Harald Janson

Mätinstrument

En central fråga i effektutvärderingar är hur man ska mäta de fenomen som tänks påverkas av interventionen. Med mätning menar man i vidaste bemärkelse att kunna beskriva aspekter av mätobjekt eller händelser i termer av tal eller kategorier (”the assignment of numbers to aspects of objects or events according to one or another rule or convention”; Stevens, 1968 i Pedhazur & Schmelkin, 1991). Kvantitativ vetenskap är beroende av mätningar av de fenomen man studerar, och genom historien har begreppsapparater och traditioner för mätmetodik utvecklats i anslutning till varje vetenskap – till exempel sociologi, medicin, pedagogik och psykologi. Olika aspekter har betonats inom olika vetenskaper; till exempel har sociologer bland annat intresserat sig för hur man kan få korrekta svar om sakförhållanden i stora undersökningar (Fowler, 2009) medan psykologer har ägnat sig åt bland annat hur man kan mäta storheter som personlighetsdrag och preferenser som är icke-observerbara i meningen att inte vara sakförhållanden som går att direkt observera eller verifiera (Weiner & Greene, 2008), och pedagoger och psykologer har utformat metodik för att testa prestationer och färdigheter (Kaplan & Saccuzzo, 2009).

Ett mätinstrument är varje form av systematisk metod för att försöka mäta en egenskap eller variabel. Egenskaper hos mätinstrument bedöms traditionellt i termer av reliabilitet (noggrannhet) och validitet (giltighet). Det här kapitlet ger en översikt över teoretiska

och praktiska överväganden inför val av mätinstrument och utformningen av datainsamlingen för en effektutvärdering.

Önskvärda egenskaper för datainsamlingen som helhet i en effektutvärderingsstudie

Det viktigaste målet med datainsamlingen i en effektutvärderingsstudie är att få goda mått på utfallsvariablerna och andra variabler av intresse i projektet. Samtidigt är det av största betydelse för att motivera och behålla deltagare i projektet att den totala tiden och insatsen som krävs av deltagarna är rimlig och att datainsamlingen är användarvänlig, attraktiv och ändamålsenlig.

Kort och användarvänlig

Forskare vill ofta inkludera många omfattande mätinstrument när de planlägger en studie. Drivkrafter för detta kan vara att kunna studera effekterna av många möjligtvis relevanta prediktorer eller mediatorer, att ta med instrument man har ett personligt intresse i ”när man ändå håller på” och rädsla för att missa någon aspekt som kan visa sig viktig. Speciellt i effektutvärderingsstudier är det av största vikt att begränsa mätningarna till det centrala och göra mätningen så kort och så användarvänlig som möjligt. För det första innebär de krav som ställs på effektutvärderingar att huvudhypotesen om effekt ska vara specificerad på förhand och gälla de centrala utfallsvariablerna; andra resultat tilläggs mindre vikt vid publicering. För det andra är det viktigt att göra allt som är möjligt för att motivera och hålla kvar deltagare i projektet med tanke på de negativa konsekvenserna av bortfall. Upplevd tids- och arbetsinsats, särskilt för den första datainsamlingen, kan spela stor roll för fortsatt intresse att delta.

Innehållet i datainsamlingen bör prioriteras som följer:

- De centrala utfallsvariablerna som antas påverkas av interventionen måste mätas med så valida, reliabla och förändringskänsliga mätinstrument som möjligt. Samma överväganden gäller för

eventuella prediktor- och moderatorvariabler som specificeras i frågeställningarna.
- Relevanta bakgrunds- och kontrollvariabler bör inkluderas med så föga krävande mått som möjligt (variabler av intresse kan vara t.ex. ålder, kön, civilstånd, familjesammansättning, utbildningsnivå, sysselsättning, språkbakgrund, eventuell minoritetsstatus och relevanta förhållanden för gruppstatus i undersökningen).
- Variabler som inte är motiverade utifrån projektets frågeställningar bör inte tas med.

Datainsamling med moderna elektroniska medier

Det var inte länge sedan nästan alla samhällsvetenskapliga forskningsdata registrerades på papper för senare dataregistrering. Var och en som har gjort detta känner till alla problem som då måste hanteras: papper kan komma bort i posten eller under andra delar av processen, sidor kan saknas i protokoll eller frågeformulär, respondenter kan hoppa över uppgifter eller hela sidor, handskriften kan vara oläslig, flera eller färre svarsalternativ än avsett kan ha kryssats för, ogiltiga svarsvärden kan ha uppgetts, hoppanvisningar kan vara felaktiga eller missvisande, korrekta hoppanvisningar kan ignoreras av respondenter, inmatning av data är tidsödande och genererar nya fel som måste kontrolleras och så vidare. Datorbaserade alternativ till datainsamling har funnits länge, men först i det 21:a århundradet har de tekniska alternativen blivit så lättillgängliga och attraktiva att helelektronisk datainsamling börjat tränga ut pappersbaserad. I dag finns inte några goda argument för att inte använda elektroniska medier för alla typer av datainsamling och registrering, såväl datainsamling där deltagaren själv bidrar med svar eller responser som data som registreras av forskarna.

Fördelarna med datorbaserad datainsamling är många. Framför allt att det kan eliminera de flesta problem med pappersmediet som räknas upp ovan. Det kan också göra att forskningen uppfattas som tidsenlig och att besvarandet av frågor blir attraktivt och användarvänligt.

Nackdelarna är framför allt två. För det första kräver det större arbetsinsats och speciella färdigheter att förbereda datainsamlingen – till exempel för att datorisera frågeformulär, registreringsprotokoll eller testprocedurer. För det andra riskerar man att data går förlorade vid tekniska problem. Nackdelen med den initiala arbetsinsatsen uppvägs mer än väl (också för ganska små forskningsstudier) av vinsten med att få data färdiga för analys direkt in i forskningsdatabasen utan de flesta av felkällorna vid pappershantering. Dataförlust genom tekniska problem måste förebyggas genom goda säkerhetssystem och rutiner, och denna risk måste vägas emot både vinsterna i form av kompletta data med elektroniska system och riskerna för dataförlust som finns också med pappersbaserade system.

I detta kapitel behandlas inte specifika tekniska lösningar för datorbaserad datainsamling, men möjligheterna är många. Frågeformulär till deltagare eller personal som är engagerade i projektet kan distribueras via internet och resultaten levereras direkt in i forskningsdatabasen (flera lösningar för detta erbjuds kommersiellt). Intervjuer, kliniska undersökningar, testning och observationsregistrering kan alla administreras och registreras direkt i en dator eller en handhållen registreringsenhet och direkt eller senare överföras till forskningsdatabasen. Deltagare eller personal kan förses med personliga elektroniska registreringsenheter för registrering eller inrapportering av till exempel checklistor eller skattningsskalor som önskas besvarade med täta eller oregelbundna intervall. Svar kan inhämtas via mobiltelefon, e-post, speciellt designade webbsidor eller applikationer som deltagarna installerar på sin egen dator eller mobiltelefon. Vid planeringen av en effektutvärdering bör man utnyttja de möjligheter som finns, men undvika alternativ som är onödiga i förhållande till frågeställningarna. Teknik ska underlätta, inte komplicera. Inkluderingen av elektroniska medier måste vidare naturligtvis vara integrerad i de etiska överväganden för forskningsstudien, och säkerheten vid registrering och databehandling av personupplysningar och andra känsliga uppgifter måste tillgodoses.

Några typer av mätningar och mätinstrument

Möjligheterna för vilka typer av instrument som kan användas för mätning av utfallsvariabler i en effektutvärderingsstudie är nästan outtömliga. Nutida forskningstradition har en stark slagsida mot frågeformulär som besvaras av deltagarna själva eller andra respondenter. Beroende på vilka variabler man önskar mäta kan det hända att också andra tillvägagångssätt kan vara aktuella. Här följer en inte uttömmande lista över olika typer av mätinstrument:

Frågeformulär

Med frågeformulär avses här alla typer av sammanställningar av frågor som besvaras skriftligt (via elektroniskt medium eller på papper) av respondenten. Detta är en billig och effektiv metod att samla upplysningar om en mångfald av fenomen. De svar som erhålls gäller respondentens egen subjektiva uppfattning. Detta är på en och samma gång metodens fördel och nackdel. Respondenten själv är i många fall den som har bäst underlag för att bedöma viktiga förhållanden. Samtidigt kan varierande språklig förståelse, kunskap om fenomenen formuläret frågar om eller självinsikt vara hinder för att få valida självrapporter. Dessa begränsningar är väl kända och ofta diskuterade i litteraturen om självrapporter (t.ex. Hunsley, 2009).

Den skriftliga formen har fördelen att innebära liten eller ingen möjlig påverkan från undersökningsledare eller intervjuare (i form av verbal och ickeverbal kommunikation när frågan ställs, och förväntningar om intervjuarens reaktion när svaret avges). Den skriftliga formen kan också vara att föredra för känsliga frågor såsom personlig ekonomi, sex, missbruk och brott.

En typ av upplysningar som i nästan alla undersökningar insamlas med hjälp av frågeformulär eller strukturerad intervju är bakgrundsdata, såsom deltagarnas utbildning, ålder, yrke, hälsa och så vidare. Det räcker med en eller några få frågor per upplysning. Man bör vara noga vid utformandet av sådana ”single items”. Helst bör frågor om vanliga demografiska variabler ställas på exakt samma sätt och med samma svarskategorier som tidigare har använts i sto-

ra befolkningsundersökningar (antingen undersökningar utförda av landets statistiska centralbyrå eller andra stora etablerade undersökningar). Då kan grupperna i effektutvärderingsstudien jämföras med en större population. Om man exempelvis vill fråga om rökning i en norsk undersökning, bör man använda den uppsättning med frågor som Statistisk sentralbyrå har använt under många år; den första frågan ska lyda ”Hender det at du røyker?” och om deltagaren svarar ja, ska nästa fråga vara ”Røyker du daglig eller av og til?” och så vidare (Statistisk sentralbyrå, 2008). Vid utformningen av ”single items” som inte kan hämtas från en tidigare undersökning är det viktigt att tänka på att frågorna är enkla, otvetydiga och har distinkta svarsalternativ. Det finns en omfattande litteratur om utformningen av befolkningsundersökningar (survey) som är relevant för utformningen av sådana frågor (t.ex. Fowler, 2009).

Standardiserade skalor

Ett specialfall av frågeformulär är standardiserade skalor, med en fackterm också kallade standardiserade inventorier. Den vanligaste formen för ett standardiserat inventorium är en instruktion som förklarar hur inventoriet ska besvaras, följt av en rad frågor som alla har samma svarsformat (t.ex. Ja/Nej, en flergradig skala av så kallat Likert-format, eller en frekvensskala). Svaren på ett antal frågor summeras (eller sammanräknas enligt en annan bestämd formel). Standardiserade inventorier är det vanligaste sättet att erhålla självrapporter om många psykologiska variabler. Ordet ”standardiserad” används i olika betydelser i det här sammanhanget. I den mest grundläggande betydelsen avses bara att innehållet är fastställt och utprövat. Ibland avses den mer specifika betydelsen som tas upp senare i detta kapitel.

Intervju

Svar på frågor kan inhämtas via frågeformulär eller intervju. Man bör välja den metod som passar bäst i det enskilda fallet, och beroende på frågeställningarna kan det vara lämpligt att använda både

intervju och frågeformulär i samma datainsamling. Till intervjumetodens fördelar hör att deltagare kan uppskatta det personliga bemötandet. Många personer föredrar att samtala med en person en stund framför att få ett frågeformulär eller en internetbaserad enkät att besvara, även om intervjun tar lite längre tid. Med personlig kontakt kan deltagare lättare känna sig delaktiga i projektet och känna sig mer motiverade att delta. Bland annat av dessa skäl har intervjuundersökningar med personlig uppföljning ofta större deltagande och mindre bortfall över tid än undersökningar som enbart samlar in upplysningar skriftligt.

När det är komplexa frågor eller komplicerade hoppinstruktioner kan en intervju ge mer kompletta och riktiga svar. När svaren ska sammanfatta eller kategorisera komplicerade eller omfattande förhållanden kan intervjuare göra detta mer tillförlitligt (och tillförlitligheten i t.ex. kategoriseringarna kan och bör prövas och redovisas empiriskt). Några typer av mätning kräver intervju på grund av innehållets natur, till exempel psykiatrisk diagnostisk och frågeställningar av kvalitativ art (Kvale, 2006). Telefonintervjuer, som bör hållas korta, har med mobiltelefonens allmänna utbredning blivit en tilltalande möjlighet att hålla tät kontakt med deltagare och erhålla frekventa mätningar av utfallsvariabler (t.ex. dagliga nivåer av ångest eller symtom eller barns problembeteende; Chamberlain & Reid, 1987).

Diagnostisk bedömning

Om frågeställningarna för effektutvärderingsstudien gäller (psykiatriska) diagnoser, måste mätningen omfatta en diagnostisk bedömning – vanligtvis en strukturerad, eventuellt datorstödd, intervju, som måste utföras av personal som har adekvat utbildning och träning i den specifika bedömningsmetoden. Reliabiliteten i den diagnostiska bedömningen bör redovisas vid publicering av resultaten.

Testning

Bland annat variabler som gäller förmågor, begåvning, prestation,

kognitiv funktion, kunskaper och färdigheter mäts med fördel med hjälp av standardiserade test. Testning innebär att respondenten får en uppsättning standardiserade uppgifter att lösa, som poängsätts efter fastställda kriterier. När testning är aktuellt i en effektutvärdering väljs test som är standardiserade och normerade. Testning ska utföras av personal med relevant utbildning och adekvat träning i den specifika testmetoden.

Observation

När utfallsvariabeln är en persons beteende eller samspelet mellan två eller flera personer, kan observation av beteendet eller samspelet vara en möjlig mätmetod. Observation har fördelen att vara en objektiv informationskälla (i bemärkelsen oberoende av deltagarnas egna uppfattningar eller bedömningar). Denna fördel är störst när intresset gäller en variabel som deltagaren har svårt att rapportera om själv. Forskning på olika områden visar varierande (oftare låg än hög) överensstämmelse mellan självrapporter och observationer och varierande validitet för den ena eller andra typen av upplysningar.

Observationsmetoder utgår antingen från naturligt beteende (t.ex. i hem- eller skolmiljö) eller från en strukturerad situation (deltagaren eller deltagarna får utföra vissa uppgifter i en standardiserad miljö, t.ex. ett laboratorium). Det är viktigt att välja miljö och tidsrymd för en observation som kan antas ge rimlig variation i det beteende man önskar mäta. Strukturerade situationer kräver betydligt kortare tid och ger data som är lättare att jämföra över personer, men observation av naturligt beteende kan vara mer giltig för beteende i verkligheten. Registreringen av observationerna kan antingen genomföras i efterhand utifrån video- eller ljudinspelning eller direkt vid observationen (t.ex. med en handhållen elektronisk registreringsenhet). Direktregistrering kräver väsentligen mindre arbetsinsats och är särskilt effektiv om det är få och lättobserverade beteendekategorier som ska registreras. En annan fördel med direktregistrering är att man inte registrerar ovidkommande perso-

ner (såsom t.ex. de andra barnen i en daghemsgrupp). Å andra sidan ger registrering från inspelat material mycket större möjligheter för mångfacetterad och detaljerad registrering (kodning eller skattning) också av subtila fenomen.

Till en effektutvärderingsstudie ska man välja observations- och registreringsprocedurer som tidigare är utprövade och validerade. Den som överväger att använda en observationsmetod hänvisas till litteraturen om observationsmetodik generellt (Bakeman & Gottman, 1997) och till specifik litteratur som handlar om observationsmetodik för den utfallsvariabel det gäller. Oavsett vilken specifik observationsmetod som väljs är det viktigt med god träning och tät uppföljning av dem som genomför observationerna. Om registrering sker från inspelat material är det också viktigt med träning och uppföljning av dem som kodar eller skattar materialet. Ett speciellt centralt moment för alla typer av observationer är att fortlöpande undersöka interbedömarreliabiliteten (vid direktobservation innebär detta att två eller flera direktobservatörer måste vara med vid en del av observationerna). Interbedömarreliabiliteten ska också kunna dokumenteras i publikationer. Den som tidigare inte har använt observationsmetodik bör ta direktkontakt med forskare som har använt den specifika metoden av intresse för att få den nödvändiga överföringen av explicit och implicit kunskap och know-how.

Registerdata

När mottagarna av interventionen befinner sig i en organisation (t.ex. sjukvården, skolan, socialtjänsten eller en arbetsgivare) kan det hända att data och händelser som rutinmässigt registreras i systemet kan användas som forskningsdata i effektutvärderingsstudien, exempelvis patientstatistik, patientjournaler, skolbetyg, närvaroregistrering, registrering av andra skolhändelser, flextidsregistreringar och arbetsprestationer. Registerdata används inte alltid till sin fulla potential på grund av olika hinder förbundna med bruk av denna typ av data. Det kan vara tidsödande och svårt att få registrerade data ut ur systemet i en användbar form. De etiska dilemman som

är förbundna med att använda systemdata i forskning måste tas på allvar; registerdata kan omfatta känsliga personupplysningar. Att inte ta i bruk redan insamlade data ökar å andra sidan belastningen på respondenterna. En viktig fråga vid övervägandet om att använda registerdata gäller datakvaliteten, som inte utan vidare kan antas vara hög utan måste värderas i varje enskilt fall. Om en intervention förankras i en stor organisation (t.ex. ett skolsystem) och utövas över längre tid kan det vara ändamålsenligt att införa registreringar av interventionsvariabler i de existerande systemen i organisationen, till bruk både för effektutvärderingen och för systemets egna behov.

Några andra datainsamlingsmöjligheter

Det finns flera typer av instrument än de som har nämnts ovan. Dagboksmetoder och periodiska rapporter kan användas för att registrera variabler eller händelser med dagliga eller periodiska intervall. Automatisk registrering med elektroniska registreringsenheter som deltagarna bär med sig kan användas för att registrera variabler som puls, blodtryck, fysisk aktivitet eller rörelser i geografin (med GPS). Provtagning kan vara aktuell när frågeställningarna gäller variabler som kan antas ge meningsfull variation i till exempel nivåer av stresshormoner. Insamling av genetiskt material (t.ex. genom blodprov eller skrapprov) kan vara aktuell om frågeställningarna gäller individuell genetisk variation (genetisk specialkompetens krävs för sådan forskning, som också medför specifika etiska dilemman och tekniska utmaningar). Indirekta mått (”unobtrusive measures”) omfattar att använda ”spår” av deltagarnas aktivitet (såsom kontorsavfall eller producerade dokument) till att registrera variabler av intresse (Shaughnessy, Zechmeister & Zechmeister, 2009; Webb, 1966). Det finns många typer av datainsamling som kan användas i effektutvärderingar och fördelarna eller nackdelarna med olika metoder måste bedömas i varje enskilt fall. Listan över önskvärda egenskaper hos mätinstrument som följer kan användas för att komma fram till ett lämpligt mätinstrument i varje enskilt fall.

Önskvärda egenskaper hos mätinstrument för effektstudier

Vid valet av mätinstrument i en effektutvärdering är det flera egenskaper hos mätinstrumentet som bör tillgodoses. Framställningen nedanför listar viktiga aspekter med de viktigaste först. Övervägandena gäller i första hand val av mätinstrument för de specifika utfallsvariabler som en intervention är riktad mot.

Instrumentet mäter rätt variabel (innehållsvaliditet)

Välj mätinstrument som så nära och uttömmande som möjligt mäter precis de utfallsvariabler som interventionen avser att förändra (och inga andra). Hur självklart detta än kan låta, är det inte ovanligt att det är dålig överensstämmelse mellan det en forskare önskar att studera och de mätinstrument som ingår i effektutvärderingen. Exempel på dålig överensstämmelse kan vara att använda symtomchecklistor för att mäta förbättring av psykisk hälsa (ett begrepp som omfattar mer än symtom på ohälsa), eller att använda föräldrars självrapporter om upplevd stress i föräldrarollen som mått på ökade föräldrafärdigheter. Valet av mätinstrument bör utgå från en noggrann förståelse av det som interventionen avser att påverka. Om en intervention exempelvis avser att minska barns aggressiva beteende hemma och i skolan genom att öka positiva föräldrabeteenden, behövs mätinstrument som mäter positiva föräldrabeteenden (den variabel som manipuleras direkt av interventionen) samt barns aggressiva beteenden hemma och i skolan (de variabler som interventionen avser att förändra).

Begreppen som mäts av mätinstrumentet bör vara operationaliserade på ett sätt som är konsistent med definitionerna av och antagandena om begreppen i interventionsmodellen. Man bör till exempel välja mått på observerbart beteende om interventionen avser att förändra beteende och mått på attityder om interventionen avser att förändra attityder. Det är också viktigt att tillgodose en mer detaljerad överensstämmelse mellan definitionen av utfallsvariabeln

i interventionen och innehållet i mätinstrumentet. Mätningen av utfallsvariablerna bör vara uttömmande, och mätningen bör vara avgränsad till relevanta aspekter av utfallsvariablerna. Om till exempel den intervention som ska utvärderas definierar positiva föräldrafärdigheter som klara besked, uppmuntran av önskvärt beteende och ignorering av negativt beteende, bör alla dessa aspekter av föräldrafärdigheter mätas, men inte andra. I litteraturen kallas denna matchning av innehåll i mätinstrument till teoretiskt begrepp för innehållsvaliditet, och expertbedömningar är det vanligaste sättet att avgöra om innehållsvaliditeten är tillfredsställd (Messick, 1981).

Empiriskt stöd för instrumentets begreppsvaliditet och kriterievaliditet

Med validitet, ett begrepp från den psykologiska testtraditionen, menas empiriskt stöd för att en mätmetod mäter det den avser att mäta. Validitet är svårt att uppskatta och mer komplicerat än vad man gärna tror (Messick, 1981). Stöd för ett tests validitet har traditionellt delats in i innehållsvaliditet (behandlat ovan), kriterievaliditet (att mätningen överensstämmer med andra kriterier på det man önskar mäta), begreppsvaliditet (att mätningarna har förväntade samband med varandra och med mätningar på andra begrepp) och ibland ”face validity” (att respondenter eller användare av instrumentet accepterar det som ett mått på ett visst begrepp utifrån innehållet, se nedan). Nyare formuleringar betonar begreppsvaliditetens överordnade betydelse och utökar validitetsbegreppet med aspekter av mätningars konsekvenser i samhället (t.ex. hur mätmetoder stämmer överens med kulturella värderingar, eller vilka de sociala konsekvenserna av mätningar är; Messick, 1989). Det är viktigt att inse att ett mätinstruments validitet inte är en antingen-eller-fråga som kan fastslås en gång för alla, utan att det kan finnas mer eller mindre, och starkare eller svagare stöd för instrumentets validitet. Det är också viktigt att ett mätinstruments validitet är specifikt för ett användningsområde och en population. Ett bestämt självrapportin-

ventorium skulle till exempel kunna ha giltighet för att identifiera självskadebeteende bland tonårsflickor i en skandinavisk oselekterad population, men inte passa för pojkar eller minoritetsgrupper, och heller inte vara användbart för att skilja på grader av självskadebeteende i en klinisk population.

I vidaste bemärkelse gäller begreppsvaliditet frågan om att kunna göra sannolikt att ett mätinstrument mäter just det begrepp det är avsett att mäta och inte något annat. Empiriskt stöd för detta utgörs av publicerade studier som visar att ett bestämt mått har ett empiriskt samband med andra mått på relaterade variabler (ibland kallat konvergent validitet), men inte har samband med mått på variabler som antas vara orelaterade (divergent validitet). Ett empiriskt samband betyder ett samband av en rimlig magnitud mätt med ett passande sambandsmått (t.ex. korrelation, medelvärdesskillnad, oddskvot, skillnad i proportioner identifierade individer mellan grupper). För att visa att ett mätinstrument är utformat specifikt för att mäta individers benägenhet att använda våld under stress är det alltså inte nog att mätinstrumentet uppvisar samband med andra mått på samma begrepp, utan det behövs också empiriskt stöd för att instrumentet inte bara mäter våldsbenägenhet, impulsivitet eller stresskänslighet i största allmänhet.

För självrapportinventorier eller andra mätinstrument som innehåller flera delskalor är studier av faktorstrukturen som visar att sambanden mellan de olika delskalorna eller variablerna är rimliga i förhållande till teoretiska förväntningar också relevant stöd för begreppsvaliditeten. Stödet för mätinstrumentets begreppsvaliditet är mer övertygande ju mer användningsområdet i den publicerade studien liknar den tilltänkta utvärderingsstudien, och ju mer undersökningsgrupperna liknar varandra. Om syftet till exempel är att studera en viss interventions effekt på lättare depression bland norska småbarnsmödrar, utgör publicerade resultat om det tänkta mätinstrumentets samband med andra mått på lättare depression bland mödrar i Skandinavien starkare stöd än publikationer från andra länder eller publikationer om kvinnor som inte har små barn,

om män, om personer med allvarlig depression eller andra psykiatriska problem och så vidare.

Vid val av ett mätinstrument för utfallsvariabler i en effektutvärderingsstudie är det speciellt viktigt att dokumentera i vilken grad det finns relevant empiriskt stöd för kriterievaliditet för ett mätinstrument i en liknande användning som den planerade. Kriterievaliditet avser att det är ett empiriskt samband mellan mätinstrumentet och ett kriterium (ett ”facit”) på det som mätinstrumentet avser att mäta. Exempel på vad som kan anses vara sådana kriterier kan vara grupper med kända egenskaper, diagnostiska bedömningar, expertvärderingar, framgång eller misslyckande i utbildning, yrke eller behandling och så vidare. Om syftet är att bedöma grupptillhörighet, till exempel om individer i en undersökningsgrupp har en bestämd diagnos, uppfyller ett beteendekriterium eller tillhör en riskgrupp, är det viktigt att undersöka mätinstrumentets sensitivitet och specificitet. Med sensitivitet menas i vilken grad instrumentet förmår att identifiera personer som enligt ett referenskriterium tillhör målgruppen, och med specificitet menas i vilken grad instrumentet inte felidentifierar personer som enligt referenskriteriet inte tillhör målgruppen (Simon & Boring, 1990).

Om det bara finns ett svagt stöd för ett mätinstruments begrepps- eller kriterievaliditet för en användning lik den planerade finns det anledning till oro. Även om resultaten är signifikanta och i förväntad riktning, kommer det att behövas underbyggda argument som talar för att det som skulle mätas faktiskt har blivit mätt. Om signifikanta resultat uteblir kan det bero på att interventionen inte fungerade eller på att de använda instrumenten var olämpliga för att mäta utfallsvariabeln.

Empiriskt stöd för mätinstrumentets reliabilitet

Reliabilitet, eller tillförlitlighet, är ett begrepp från den psykologiska testtraditionen. Med reliabilitet menas i vidaste bemärkelse i vilken grad variationen i det kvantitativa resultatet av en mätning reflekterar variationen i den variabel man önskar att mäta. Detta är i

motsats till oönskad variation (ibland kallat mätfel) som kan bero på till exempel fluktuation i mätningar över tid, skillnader mellan olika mätinstrument eller deluppgifter som antas mäta samma sak (item/frågor) eller observatörs- eller kodarvariation.[1] Reliabilitet är inte en egenskap hos ett mätinstrument, utan en egenskap hos resultat (poäng eller andra kvantitativa resultat) som har erhållits med hjälp av instrumentet i en bestämd population. Reliabiliteten för ett instrument kan vara hög i en population och låg i en annan. Till exempel kan reliabiliteten hos kliniska skalor vara högre i kliniska populationer än i normalpopulationer. Reliabiliteten för en ny standardisering eller översättning av ett instrument kan vara annorlunda än för en tidigare version. Liksom för validitet kan man därför inte fastslå ett instruments reliabilitet en gång för alla; det kan finnas mer eller mindre relevant stöd för mätresultats reliabilitet i en population som liknar den man önskar att undersöka och för en liknande användning som man planerar. Reliabilitetestimat från relevanta populationer är grundläggande information om ett mätinstrument. Reliabilitet är en förutsättning för validitet. Reliabiliteten i mätresultat kan därmed vara hög utan att det finns validitet (d.v.s. mätningen kan vara noggrann, men gäller inte det man avser att mäta), men validiteten kan i princip inte vara högre än reliabiliteten. Det är ett utbrett missförstånd att denna princip kan förstås bokstavligt så att till exempel en låg alfakoefficient måste innebära att en skala är invalid. Detta är felaktigt eftersom alfa och andra metoder att uppskatta reliabilitet alla har begränsningar och kan vara missvisande som mått på reliabiliteten i en egentligare betydelse. Reliabilitet kan estimeras med flera olika metoder, som tar hänsyn till olika typer av oönskad variation. De vanligaste reliabilitetsestimaten är inre konsistens, split-half och parallelltestreliabilitet, test-retest-reliabilitet och interbedömarreliabilitet.

1 För en utförligare framställning av reliabilitetsproblematiken, se t.ex. Cronbach, Nanda, Gleser & Rajaratnam (1972) och Shavelson & Webb (1991).

Inre konsistens

När ett mätinstrument, till exempel ett inventorium, en checklista eller ett psykologiskt test, är uppbyggt av flera olika deluppgifter eller frågor som förutsätts mäta samma underliggande variabel, är det av intresse att uppskatta hur mycket av variationen i mätresultatet som kommer av det som deluppgifterna mäter gemensamt (och som kan antas vara variation i den underliggande variabeln), till skillnad från variation som är unik för varje deluppgift. Inre konsistens, som typiskt mäts med koefficienten Cronbachs alfa (Cronbach, 1951, se också Streiner, 2003a) är ett uttryck för samvariation i poänger på deluppgifterna, mer specifikt kvoten mellan gemensam variation i poäng över deluppgifter och hela variationen i mätresultatet. I mer vardagliga termer kan man säga att en hög inre konsistens tyder på att deluppgifterna ”mäter detsamma”. Hög inre konsistens är direkt beroende av antalet deluppgifter eller testlängden: Ju fler deluppgifter, desto högre inre konsistens. Ett annat förhållande som påverkar inre konsistens är vidden av begreppet som mätinstrumentet ska mäta: den inre konsistensen blir lägre för vidare begrepp (d.v.s. det behövs flera frågor eller item för att uppnå samma inre konsistens för ett vitt begrepp som social kompetens än för ett homogent begrepp som examensångest). Inre konsistens är tekniskt lätt att beräkna och finns normalt angivet i manualer eller publikationer om mätinstrument. Välkonstruerade inventorier, checklistor och psykologiska test bör kunna dokumentera en hög inre konsistens i relevanta populationer. Vad som kan anses vara högt måste ses i sammanhang med instrumentets natur och det man önskar mäta. En inre konsistens på 0,80 anses högt i de flesta sammanhang, men också mätinstrument med lägre inre konsistens kan försvaras beroende på omständigheterna och designen. Det är viktigt att inse att inre konsistens estimerar en begränsad aspekt av reliabilitet (att poängen på uppgifterna i testet samvarierar) och inte är bevis för att mätinstrumentet är reliabelt i en vidare bemärkelse (t.ex. över tid, situationer eller bedömare). Många har också varnat för att förlita sig alltför mycket på inre konsistens som mått på ett mätinstruments

reliabilitet och påpekat problem med detta (Schmidt, Le & Ilies, 2003; Streiner, 2003b). Vid val av mätinstrument för utfallsvariabler i en effektutvärdering bör man sträva efter att välja instrument med hög inre konsistens i tidigare mätningar i relevanta populationer. De främsta skälen för att eftersträva reliabla mätinstrument är att dessa ger större power och bättre möjlighet att påvisa effekter, men det finns också pragmatiska skäl. En onyanserad förståelse av inre konsistens är utbredd, och det förekommer att författare möter kategoriska invändningar mot mätinstrument som uppvisar en alfa på mindre än 0,70 vid publicering i vetenskapliga tidskrifter. När det är relevant bör den inre konsistensen i mätningen i effektutvärderingsstudien redovisas vid publiceringen.

Split-half-reliabilitet och parallelltestreliabilitet

Split-half-reliabilitet estimerar detsamma som inre konsistens men på ett begränsat sätt, nämligen att dela in deluppgifterna i två ”deltest” (t.ex. udda och jämna deluppgifter) och beräkna sambandet mellan mätningarna av deltesten. Split-half-reliabilitet redovisas sällan nuförtiden och kan anses vara ekvivalent med inre konsistens. Ett annat relaterat sätt att estimera hur mycket av variationen i mätresultat som beror på valet av specifika deluppgifter är parallelltestreliabilitet. Det går ut på att konstruera två olika separata test utifrån samma principer och mäta hur stor överensstämmelse det är mellan mätresultaten från de båda testen. Hög parallelltestreliabilitet utgör ett starkare stöd för mätresultatens oberoende av variationer i deluppgifter än inre konsistens, men undersöks sällan på grund av kostnaderna.

Test-retest-reliabilitet

Denna metod uppskattar hur mycket av variationen i mätresultatet som inte beror på oönskad fluktuation över tid. Två mätningar genomförs med ett passande tidsintervall emellan (t.ex. två veckor); så långt att respondenterna inte har mätinstrumentet i färskt minne och tidsrelaterade fluktueringar kan ha antagits inträffa, men

inte så lång tid att respondenterna kan antas ha förändrat sig betydligt med avseende på den variabel mätningen avser. Sambandet mellan mätningarna utgör estimatet av test-retest-reliabiliteten. Ett högt estimat av test-retest-reliabiliteten visar att variationen i mätresultat i liten grad beror på dagsformen eller andra faktorer som kan orsaka variation i mätningen över kortare tidsintervall. Om det finns betydande träningseffekter (d.v.s. respondenterna ”lär sig” ett mätinstrument vid den första mätningen) är det problematiskt att uppskatta test-retest-reliabiliteten; för typiska inventorier och checklistor kan man anta att träningseffekter inte är betydande. Test-retest-reliabilitet redovisas mer sällan än inre konsistens på grund av kostnaderna för att genomföra test-retest-undersökningar. Om det finns tidigare resultat som visar hög test-retest-reliabilitet i relevanta populationer i tillägg till hög inre konsistens för ett mätinstrument, är det i allmänhet starkare stöd för ett mätinstruments reliabilitet i en vidare bemärkelse än om det bara finns resultat som visar hög inre konsistens. Hög test-retest-reliabilitet kan naturligtvis inte förväntas för testinstrument som mäter variabler som antas förändras snabbt. Specifikt för effektutvärderingsstudier finns det därmed en motsättning mellan förändringskänslighet i ett mätinstrument och test-retest-reliabilitet: instrument som är känsliga för förändring kan ha lägre test-retest-reliabilitet. Det förväntas inte att man ska kunna dokumentera test-retest-resultat från den egna effektutvärderingsstudien (designen medger normalt inte att undersöka detta).

Interbedömarreliabilitet

Denna metod, också kallad interbedömaröverensstämmelse, gäller en begränsad men viktig aspekt av reliabilitet, nämligen i hur stor utsträckning två bedömare (kodare, skattare, intervjuare, testledare osv.) kan göra samma kvantitativa bedömning. Interbedömarreliabilitet estimeras som sambandet mellan mätningar gjorda av två olika bedömare, ofta med en intraklasskorrelation (ICC; Shrout & Fleiss, 1979) för kontinuerliga variabler och ett mått i kappa-famil-

jen för kategoriska variabler (Janson & Olsson, 2001; 2004; Landis & Koch, 1977). För alla mätmetoder där mätresultaten är beroende av bedömare är interbedömarreliabilitet en viktig aspekt att bedöma. Detta gäller bland annat kodning eller skattning av observationer och kvalitativa svar, poängsättning av testresultat, diagnoser och skattning eller klassificering av respondentvariabler av intervjuaren på en skattningsskala. Hög interbedömarreliabilitet är beroende av precisa definitioner och kriterier för bedömningen, observatörernas utbildning och erfarenhet samt träning och samträning i observationsmetoden. Vad som är tillräcklig eller hög interbedömarreliabilitet beror i hög grad på svårigheten i bedömningsuppgiften (högre överensstämmelse förväntas för bedömningar av enkla aspekter såsom om ett observerbart beteende är närvarande eller inte, men ju komplexare och ju mer omfattande bedömningen är, desto lägre kan överensstämmelsen förväntas vara). Vid val av en mätmetod som tar i bruk bedömare av något slag är det viktigt att dokumentera den tidigare interbedömaröverensstämmelsen i liknande användningar och populationer, men det är också nödvändigt att säkra god interbedömarreliabilitet i effektutvärderingsstudien och kunna dokumentera denna vid publicering.

Estimat av relevanta aspekter av reliabiliteten i mätningarna är viktig dokumentation om mätnoggrannheten i en effektutvärderingsstudie; i någon mån kan man också använda reliabilitetsestimat för att estimera interventionseffekter justerat för mätnoggrannhet.

Instrumentet är känsligt för förändring

Effektutvärderingar av interventioner fokuserar vanligtvis på att mäta förändringen i utfallsvariabeln (eller variablerna) som kan hänföras till interventionen. En vanlig forskningsdesign syftar till att studera skillnaden mellan mätning av utfallsvariabeln före och efter interventionen i en grupp som har tagit del av interventionen, och jämföra med skillnaden mellan mätning av utfallsvariabeln före och efter i en kontrollgrupp, men en mångfald av forskningsdesigner kan användas för att studera förändring i utfallsvariabeln. Det som

blir speciellt viktigt vid val av mätinstrument i detta perspektiv är att mätinstrumentet kan fånga upp förändringen som har skett över det tidsrum när interventionen har utförts. Mätinstrument skiljer sig med avseende på hur känsliga de är för förändring, oberoende av om instrumenten har utmärkta egenskaper för övrigt.

Helst önskar man att det finns tidigare forskning (t.ex. tidigare liknande effektutvärderingar) som visar att ett tänkt mätinstrument har kunnat fånga upp en sådan förändring som förväntas. Om sådana forskningsresultat inte finns, är det extra viktigt att göra en värdering av mätinstrumentets möjlighet att göra detta. Frågor om och bedömningar av utfallsvariabeln som ingår i mätinstrumentet bör gälla nuläget eller ett passande kort tidsrum (t.ex. sista dygnet, sista veckan, sista två veckorna) så att eftermätningen inte omfattar tiden före eller under interventionen. Svars- eller registreringsformatet bör vara tillräckligt fingraderat för att fånga upp en liten men betydelsefull förändring (t.ex. en 5- eller 7-gradig skala med verbalt definierade kategorier, eller en frekvensskala som ger möjlighet att nyanserat beskriva hur ofta ett beteende förekommer). Svarsformat som ”stämmer inte – stämmer i någon grad – stämmer precis” eller ”aldrig – ibland – ofta” kan vara för grovmaskiga för att fånga upp betydelsefulla förändringar som har ägt rum. Att mätinstrumentet för utfallsvariabeln kan förväntas ha en tillräckligt hög reliabilitet för att kunna fånga upp en förändring av den storlek som förväntas är viktigt att tillgodose. Olika informanter kan vidare också vara olika känsliga för att fånga upp förändring (den som själv har varit med om interventionen lägger gärna märke till förändringen först, närstående såsom kamrater och familj därnäst, medan t.ex. chefer eller lärare kan lägga märke till förändringen först efter en längre tid).

Om designen omfattar flera mättillfällen är det viktigt att mätningarna av utfallsvariabeln vid de olika mättillfällena sker på exakt samma sätt. Frågor och svarsformat får inte förändras. Dessutom bör inte ens minimala förändringar ske i till exempel presentationsformat, inramning eller plats för datainsamlingen eller intervjun, frågors ordningsföljd, innehåll, layout, rubriker, illustrationer, grafik

eller liknande i frågeformulär eller på skärmbilder. Också till synes triviala förändringar kan påverka respondenters svar och därmed bli en felkälla till mätningen av förändring av utfallsvariabeln (Biemer, Groves, Lyberg, Mathiowetz & Sudman, 1991). Ett undantag från kravet på exakt lika mätningar vid olika tidspunkter är långtidsuppföljningar, där man ibland måste byta ut instrumenten med hänsyn till deltagarnas ökande ålder för att undgå tak- och botteneffekter (t.ex. om man mäter barns läsförmåga över en tidsrymd av flera år).

Att mäta förändring innebär alltså att mäta utfallsvariabeln vid två eller flera tidpunkter och jämföra mätningarna. Det är något annat att ställa frågor om förändring vid en tidpunkt (t.ex. eftermätningen). Det är viktigt att inse att de svar man då får är respondentens *uppfattning* om förändring vid den tidpunkten frågan besvaras, som förutom att återspegla minnet av situationen vid förmätningen kan påverkas av uppfattningen om nuläget, förhållandet till den som har administrerat interventionen och andra faktorer. Sådana uppfattningar kan vara värdefulla i sig, och utvärderingsstudiens frågeställningar kan omfatta studiet av dem liksom av till exempel respondentuppfattningar om vad som var verksamt i interventionen. Retrospektiva uppfattningar av förändring är dock inte tillfredsställande som det huvudsakliga måttet på förändring av utfallsvariabeln som resultat av interventionen.

Mätinstrumentet är standardiserat och normerat

Mätinstrumentet till en effektutvärderingsstudie bör vara utprövat, standardiserat och gärna normerat i en grupp från en relevant population. Med standardisering menas här att proceduren eller administrationen av mätmetoden är väldefinierad inklusive instruktioner, träning av intervjuare, testledare, observatörer eller liknande. Standardiseringen innebär att alla relevanta detaljer för att kunna utföra mätningen på det avsedda sättet är skriftliggjorda i en manual eller liknande. I standardisering ingår språkanpassning, inkluderat att översätta och språkgranska mätinstrument som är skrivna eller utformade på ett annat språk och att tillgodose att språkligt

material är förståeligt för personer med de läs- och skrivfärdigheter som förväntas i målgruppen. En relaterad fråga är kulturell anpassning till en tänkt målgrupp – standardiseringen innebär att tillgodose att en mätmetod är relevant och acceptabel för den tänkta målpopulationen.

Normering innebär att det finns insamlade referensdata för en undersökningsgrupp från en relevant population, så att ett mätvärde för en individ kan jämföras med dessa referensdata för att estimera hur individen förhåller sig i relation till den relevanta populationen. För en liten effektutvärderingsstudie där ingen av grupperna är populationsbaserad är det en stor fördel om mätinstrumenten för utfallsvariablerna är normerade i en relevant population, eftersom individ- och gruppvärden i studien kan förankras i referensvärdena från populationsstudien. Ju större undersökningsgruppen är, och ju mer en av grupperna (t.ex. kontrollgruppen) liknar en väldefinierad population, desto mindre är behovet att kunna jämföra med referensdata.

Instrumentet är etablerat

Finns det relevanta mätinstrument som är ofta använda och välrenommerade bland forskare och relevanta yrkesgrupper inom fältet som interventionen gäller, bör man för en effektutvärderingsstudie alltid välja dessa framför oprövade eller mindre använda instrument. Att använda etablerade mätinstrument har många fördelar. Kommunicerbarhet, det vill säga att läsarna eller användarna av studien förstår vad resultaten betyder när de har kännedom om mätinstrumentet, är en viktig aspekt att ta hänsyn till. Resultatens trovärdighet behöver underbyggas mindre i en publikation när ett etablerat mätinstrument har använts, och det är möjligt att det är lättare att få resultaten publicerade. För ofta använda och välrenommerade instrument finns det ofta relevant stöd för validitet och reliabilitet.

Samtidigt finns det omständigheter som kan tala emot att använda etablerade instrument och man bör göra en noggrann bedömning av för- och nackdelar för den specifika tänkta användningen. Starkt

stöd för ett etablerat instruments kriterievaliditet kan till exempel behöva avvägas mot bättre innehållsvaliditet hos ett mindre etablerat instrument. Ett instrument blir etablerat genom att användas länge i forskning, men att instrumentet har funnits länge kan samtidigt betyda att det som instrumentet mäter, eller sättet det mäter det på, börjar stå långt från kunskapsutvecklingen på området. En del etablerade instrument är också dyrbara att använda.

Mätinstrumentet ligger i forskningsfronten

Mätinstrument är färskvara. Man bör välja mätinstrument som återspeglar aktuell teori och förståelse inom relevanta forskningsfält. Detta är inte minst viktigt att ta ställning till ifråga om gamla och etablerade instrument. Även sådana når slutet av sin levnadstid någon gång, och även om det finns starkt stöd för ett bestämt instruments reliabilitet och validitet ackumulerad över lång tid, kommer det till en punkt när fördelarna med ett senare utvecklat instruments bättre anpassning till moderna begrepp och teorier i forskningsfältet kan överväga. Forskare bör inte bidra till att kunskapsutvecklingen inom ett forskningsfält hämmas genom slentrianmässig användning av föråldrade mätinstrument. Ett äldre instrument kan underlätta publicering på grund av anseende och att det ofta används, men nyare instrument kan förhöja publicerbarheten genom högre relevans för aktuell kunskapsutveckling.

Mätinstrumentet är fritt tillgängligt för forskning eller kan användas till en rimlig kostnad

Det är viktigt att känna till om ett tänkt mätinstrument är fritt tillgängligt för forskning eller om det krävs tillstånd eller betalning för användning. Även om man kan finna mätinstrumentet i sin helhet i en artikel, bok eller på internet, kan man inte anta att instrumentet är fritt tillgängligt för forskning om inte detta är uttryckligen angivet. Upphovsmäns och copyrightinnehavares rättigheter ska respekteras, och det är inte acceptabelt att använda ett mätinstrument utan erforderlig betalning eller tillåtelse. Det kostar pengar att ut-

veckla goda mätinstrument. Att instrument används i forskning bidrar å andra sidan till att ackumulera empiriskt stöd för instrumentets reliabilitet, validitet och användbarhet. Det är därför vanligt att rättighetsinnehavare på förfrågan ger tillåtelse till användning av mätinstrument i forskning antingen gratis eller till rabatterat pris. Sådan tillåtelse bör vara skriftligt dokumenterad för att kunna göras gällande i framtiden, inte minst vid publicering. För att erhålla tillåtelse till användning av mätinstrumentet tar man kontakt med rättighetsinnehavaren, som är förlaget om instrumentet är kommersiellt publicerat, eller annars upphovsmannen (författaren).

Mätinstrumentet godtas som relevant av deltagarna ("face validity")

Effektutvärderingsforskning bygger som regel på frivilligt och informerat samtycke. Deltagarna har då tagit ställning till information om projektets syfte, vad de som medverkande ska bidra med och vad resultaten ska användas till. Om deltagarna inte förstår hur ett mätinstrument är relevant för projektet de har tackat ja till att delta i, kan det orsaka avhopp, ickesvar eller andra oönskade reaktioner. Det är viktigt att informationen om projektet är anpassad till innehållet, och det är viktigt att inte inkludera mätinstrument som är svåra att motivera för deltagarna utifrån den information de har fått om projektet. Speciellt känsligt innehåll bör man undvika att inkludera i undersökningen om det inte är välmotiverat utifrån projektets syften. Detta kan gälla bland annat våld, lagbrott, övergrepp, psykiska eller kroppsliga symtom, suicidtematik, problembeteende, sex, alkohol, droger, personlig ekonomi, skatt, politiska ställningstaganden och religion. Naturligtvis ska man inkludera mätinstrument om känsliga områden om de är relevanta för projektets frågeställningar, något som bör återspeglas tydligt i informationen om projektet till deltagarna. Närgångna undersökningsformer – till exempel psykologisk testning, diagnostisk intervju, läkarundersökning och videoinspelning av samspel – är också viktiga att värdera i relation till upplevd relevans i förhållande till undersökningens syften.

Det kan vara särskilt viktigt att tänka på relevansen i relation till eventuella oselekterade undersöknings- eller kontrollgrupper med låg problemförekomst; grupper som har sökt hjälp accepterar lättare att tillfrågas om problem av det slag de har sökt hjälp för och andra relevanta personliga förhållanden.

Om ingenting passar?

Sällan finner man mätinstrument som har alla de nu nämnda önskvärda egenskaperna utan för- och nackdelar hos olika instrument måste vägas mot varandra. Valet kan vara svårt – ska man föredra ett instrument som har god anpassning till de begrepp som interventionen antas påverka och ligger i forskningsfronten men bara har begränsat stöd för reliabilitet och validitet, eller ett etablerat instrument som sämre matchar begreppen man vill studera?

Om man skulle komma fram till att det saknas mätinstrument som är goda nog för ändamålet, står man i en svår situation. En effektutvärderingsstudie är knappast det rätta tillfället att testa ett oprövat mätinstrument. Den värsta tänkbara situationen som kan uppstå vid användning av ett oprövat mätinstrument är att det inte går att avgöra om bristen på interventionseffekter beror på att interventionen är overksam eller på att mätinstrumentet brister i reliabilitet eller validitet.

Om det saknas lämpliga instrument kan ett alternativ vara att översätta, anpassa och standardisera ett utländskt mätinstrument med önskvärda egenskaper. Översättning kräver inhämtande av tillåtelse och bör ske efter etablerade riktlinjer (Brislin, 1970). Standardiseringen av en nyöversättning bör omfatta en egen pilotstudie.

I allra värsta fall kan man överväga att skjuta effektutvärderingen på framtiden medan man utformar och standardiserar ett nytt mätinstrument. Detta innebär ett separat forskningsprojekt för att etablera de mest basala resultaten gällande reliabilitet, fördelning i en referensgrupp och validitet (t.ex. DeVellis, 2003).

För- och nackdelar med multi-informant-design

De flesta effektutvärderingar använder dem som är föremål för interventionen och/eller tänks påverkas av den som respondenter. Många studier använder också andra respondenter, till exempel behandlande personal, föräldrar, andra familjemedlemmar, arbetskamrater, skolkamrater, vänner, lärare, fritidsledare, chefer och så vidare. Olika informanter har olika perspektiv på deltagaren och målbeteendet (som kan yttra sig olikt i olika kontexter), och dessa perspektiv kan vara mer eller mindre sammanfallande. En design där flera informanter rapporterar om en utfallsvariabel kallas multi-informant-design. I allmänhet anses detta vara en styrka i en effektutvärderingsstudie. Statens beredning för medicinsk utvärdering (SBU) har till exempel som krav för omdömet ”hög studiekvalitet” av utvärderingar att de använt minst två oberoende bedömare av effekten (t.ex. lärar- och självskattning).

Olika informanter observerar och rapporterar aldrig exakt detsamma. I själva verket är det vanligast att finna bara moderata samband mellan svar från olika informantgrupper (exempelvis ungdomar, deras föräldrar och ungdomarnas lärare). När olika informanters rapporter i någon mån konvergerar, kan en kombination av respondenter fånga upp mer variation i utfallsvariabeln och ge mer reliabla mätningar av effekter. Möjligheten finns också att separera gemensamma och respondentunika effekter (t.ex. Offord, Boyle, Racine, Szatmari, Fleming, Sanford m.fl., 1996). Olika respondenters uppfattning om effekter kan också skilja sig mycket. Till exempel kan detta gälla för grupper med speciell eller komplex problematik. I sådana fall kan en multi-informant-design ge en helhetsbild av effekter som inte hade varit möjlig utifrån enbart en informants rapporter. Till exempel kan en klient efter en intervention mot missbruk vara nöjd och uppge att missbruket minskat, medan registerdata visar att klienten har varit inlagd flera gånger för avgiftning.

Multi-informant-design medför att tolkningen av resultaten blir mer komplicerad. Vid preciseringen av frågeställningarna för en

multi-informant-studie bör det specificeras klart hur multi-informant-data ska analyseras och tolkas, eventuellt med separata specifika hypoteser om utfallen för de olika informanterna. Vid val av mätinstrument för olika informanter har man att ta ställning till om samma instrument kan användas av de olika informanterna, eller om separata versioner måste användas. Några mätinstrument har separata versioner för olika respondentgrupper, såsom Achenbachs ASEBA – en familj av mätinstrument för beteendeproblem och relaterade variabler hos barn och unga (Achenbach, 2011).

Pilotering

Ett av de stora mysterierna med forskning är hur välutbildade personer kan ägna många månader åt att skriva planer och söka pengar för en studie, ägna tusentals arbetstimmar åt datainsamling och år åt dataanalys och förmedling av resultaten, men bara en eftermiddag åt att sammanställa frågeformuläret till datainsamlingen och inte ens fem minuter till att pröva ut det.

Många har vid mer än ett tillfälle bittert ångrat försummelser vid utformning av datainsamlingen som kunde ha fångats upp vid en pilotering. Verkligheten överträffar de flesta föreställningar om vad som kan gå fel. En partiell lista över redigeringsmisstag som utomordentligt lätt kunde ha eliminerats i fråga om pappers- eller databaserade frågeformulär omfattar bristfälliga instruktioner för ifyllandet, avsaknad av plats för ID, datum eller andra viktiga identifierande upplysningar, stavfel och språkliga plumpheter, duplicerade frågor eller sidor, felaktiga eller saknade rubriker, felaktig sidnumrering, text som har fallit bort, ovidkommande text som har kommit med av misstag, otydliga svarsformat (t.ex. om det ska väljas ett enda svarsalternativ eller alla som gäller) och bristfälliga eller frånvarande hoppinstruktioner. Med bara marginellt större insats kunde många problem som gäller svårförståeligt eller svårbesvarat innehåll (t.ex. dubbeltydiga eller oprecisa frågor, eller motsägande eller otydliga svarsinstruktioner) ha identifierats och avhjälpts. Mot-

svarande gäller registreringsformat och instrument som ska fyllas i av forskningsmedarbetare eller kliniker (och sådana fel kan ha förödande effekter om det är systematiska skillnader i hur personalen tolkar otydliga instruktioner).

Standardiserade instrument som ska ingå i en undersökning får inte ändras. Däremot kan det uppstå många fel vid överföringen till eller redigeringen in i datainsamlingsformatet av standardiserade instrument som kan upptäckas vid utprövning.

Pröva därför alltid ut en tänkt datainsamling, inklusive tidigare standardiserade inslag, och pröva ut alla inslag i datainsamlingen. Erfarenheten visar att all tid och möda som ägnas åt pilotering betalar sig mångfalt. Beroende på tiden som står till förfogande och datainsamlingens natur rekommenderas en eller flera av nedanstående utprövningssteg efter att du själv har sammanställt och korrekturläst datainsamlingen:

Låt kollegor läsa igenom

Den som har sammanställt datainsamlingsinstrumentet blir lätt blind för problem i det. Det första, obligatoriska, steget i utprövningen är därför att be några kollegor besvara den tänkta datainsamlingen som om de var respondenter, och gå igenom deras svar och spontana reaktioner. I ett andra steg kan du be samma kollegor om att se över utformningen av datainsamlingen med sina forskarögon och komma med andra möjliga synpunkter. En tredje möjlig teknik är att be någon eller några personer tänka högt vid besvarandet – hur de tänker när de får en viss fråga eller instruktion, och hur de kommer fram till sitt svar. Detta kan avslöja delar av datainsamlingen som är svåra att förstå.

Pröva ut datainsamlingen på ett litet bekvämlighetsurval

Att pröva ut den planerade datainsamlingen på ett litet bekvämlighetsurval (t.ex. tio personer) som har åtminstone någon likhet med målgruppen för datainsamlingen ger ett mycket bättre test av hur datainsamlingen fungerar. Detta ger till exempel möjlighet att prö-

va ut tiden datainsamlingen tar, fånga upp uppenbart oönskade reaktioner eller svarsmönster och observera ting som respondenterna spontant reagerar på, har frågor om eller har svårt att svara på. Erfarenhetsmässigt ger det mest att instruera respondenterna att svara eller handla som om de var med i undersökningen. Om instruktionen är att vara speciellt uppmärksam på ting som respondenterna uppfattar som problem, riskerar man att få många ”falska positiva” reaktioner på inslag som inte skulle visa sig vara problem i en verklig datainsamling. En avslutande intervju kan avklara om deltagarna har reagerat eller tänkt speciellt på något i datainsamlingen. För denna typ av pilotering kan det vara nödvändigt att först avklara etiska aspekter.

Utprövning i en grupp som liknar målgruppen

Att pröva ut den planerade datainsamlingen på en grupp personer som liknar målgruppen och är utvald med en genomtänkt strategi är det bästa testet av om datainsamlingen är ändamålsenlig, acceptabel och användarvänlig för målgruppen. Om pilotgruppen är någorlunda stor (t.ex. minst 30 personer) kan det lilla datamaterialet som fås granskas för oväntade svarsmönster som kan ha att göra med felaktigheter i utformningen, och eventuellt utforskas med avseende på deskriptiv statistik. I något större urval (t.ex. minst 50 personer) kan även till exempel samband mellan variabler och inre konsistens undersökas för att se om estimaten grovt överensstämmer med förväntningarna. För denna typ av pilotering är det nödvändigt att först avklara etiska aspekter.

Sammanfattning

- Målet för datainsamlingen i en effektutvärdering är att få goda mått på utfallsvariablerna och andra variabler av centralt intresse i projektet.
- Välj mätinstrument för utfallsvariablerna med omsorg. Önskvärda egenskaper för mätinstrument är att de mäter rätt variabel, att

det finns empiriskt stöd för instrumentets begreppsvaliditet och kriterievaliditet, att det finns empiriskt stöd för instrumentets reliabilitet, att instrumentet är förändringskänsligt, att instrumentet är standardiserat och normerat, etablerat, ligger i forskningsfronten, är fritt tillgängligt för forskning och lätt godtas av deltagarna. För- och nackdelar med olika möjliga instrument måste vägas mot varandra.

- Frågeformulär är en billig och effektiv metod att samla in svar från respondenter. Tänk över om det finns andra datainsamlingsmetoder (t.ex. intervju, diagnostisk bedömning, testning, observation eller registerdata) som är aktuella med tanke på de variabler du vill mäta.
- Begränsa datainsamlingen till det som är nödvändigt för frågeställningarna. Använd elektroniska medier för att få kompletta och felfria data, och för att göra datainsamlingen tidsenlig, attraktiv och användarvänlig.
- Pröva ut datainsamlingen innan den genomförs. All tid och möda som ägnas åt pilotering betalar sig mångfalt.

Fördjupningslitteratur

Kaplan, R. M. & Saccuzzo, D. P. (2009). *Psychological testing: Principles, applications, and issues* (7th ed.). Belmont, CA: Wadsworth.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Pedhazur, E. J. & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Referenser

Achenbach, T. (2011). ASEBA®: Achenbach System of Empirically Based Assessment. Hämtad 20 april 2011 från www.aseba.org.

Bakeman, R. & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). Cambridge, England: Cambridge University Press.

Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A. & Sudman, S. (Eds.) (1991). *Measurement errors in surveys.* New York: Wiley.

Brislin, R. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1,* 185–216.
Chamberlain, P. C. & Reid, J. B. (1987). Parent observation and report of child symptoms. *Behavioral Assessment, 9,* 97–109.
Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.
Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles.* New York: Wiley.
DeVellis, R.F. (2003). *Scale development: Theory and applications* (2nd ed.). Thousand Oaks, CA: Sage.
Fowler, F. J., Jr. (2009). *Survey research methods* (4th ed.). Thousand Oaks, CA: Sage.
Hunsley, J. (Ed.). (2009). Introduction to the special issue on developments in psychological measurement and assessment [Special issue]. *Canadian Psychology, 50*(3).
Janson, H. & Olsson, U. (2001). A measure of agreement for interval or nominal multivariate observations. *Educational and Psychological Measurement, 61,* 277–289.
Janson, H. & Olsson, U. (2004). A measure of agreement for interval or nominal multivariate observations by different sets of judges. *Educational and Psychological Measurement, 64,* 62–70.
Kaplan, R. M. & Saccuzzo, D. P. (2009). *Psychological testing: Principles, applications, and issues* (7th ed.). Belmont, CA: Wadsworth.
Kvale, S. (2006). *Det kvalitative forskningsintervju.* Oslo: Gyldendal akademisk.
Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.
Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. *Psychological Bulletin, 89,* 575–588.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). New York: Macmillan.
Offord, D. R., Boyle, M. H., Racine, Y., Szatmari, P., Fleming, J. E., Sanford, M., m.fl. (1996). Integrating assessment data from multiple informants. *Journal of the American Academy of Child and Adolescent Psychiatry, 35,* 1078–1085.
Pedhazur, E. J. & Schmelkin, L. P. (1991): *Measurement, design, and analysis: An integrated approach.* Hillsdale, NJ: Lawrence Erlbaum Associates.
Schmidt, F. L., Le, H. & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods, 8,* 206–224.
Shaughnessy, J. J., Zechmeister, E. B. & Zechmeister, J. S. (2009). *Research methods in psychology* (8th ed.). Boston, MA : McGraw-Hill Higher Education.
Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory: A primer.* Thousand Oaks, CA: Sage.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.
Simon, D. & Boring, J. R., III. (1990). Sensitivity, specificity, and predictive value. In H. K. Walker, W. D. Hall & J. W. Hurst (Eds.) *Clinical methods: The history, physical, and laboratory examinations.* (Chapter 6.) Boston, MA: Butterworths. Hämtad 19 april 2011 från http://www.ncbi.nlm.nih.gov/books/NBK383/
Statistisk sentralbyrå (2008). *Røyking i Norge, 2008: Jevn nedgang i andel dagligrøykere.* Hämtat 19 april 2011 från http://www.ssb.no/emner/03/01/royk/arkiv/.
Stevens, S. S. (1968). Measurement, statistics, and the schemapiric view. *Science, 161,* 849–856.
Streiner, D. L. (2003a). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80,* 99–103.
Streiner, D. L. (2003b). Being inconsistent about consistency: When coefficient alpha does and doesn't matter. *Journal of Personality Assessment, 80,* 217–222.
Webb, E. J. (1966). *Unobtrusive measures: Nonreactive research in the social sciences.* Chicago, IL: Rand McNally.
Weiner, I. B. & Greene, R. L. (Eds.) (2008). *Handbook of personality assessment.* Hoboken, NJ: Wiley.

Kapitel 6 | Tina M. Olsson

Ekonomiska analyser

Effektutvärderingar kan utformas för att besvara flera frågor, till exempel:

- Kan en intervention fungera under optimala förhållanden (eng. efficacy)?
- Fungerar interventionen i ordinarie verksamhet (eng. effectiveness)?
- För vem fungerar interventionen (moderatoranalys)?
- Vilka är interventionens för- respektive nackdelar (nettonytta)?

Effektutvärderingar kan dessutom utformas för att svara på frågan:

- Vad kostar interventionen (kostnadsanalys)?

En ekonomisk analys som görs i samband med en effektutvärdering kallas för ekonomisk utvärdering. Där jämförs olika interventioners kostnader och effekter. Den syftar till att underlätta beslutsfattandet om hur ändliga resurser ska användas på effektivast möjliga sätt. Eftersom resurserna är begränsade innebär varje beslut om att införa en intervention att något annat måste stå tillbaka. Denna uppoffring kallas för alternativkostnad och är den nytta som man måste avstå från till följd av de val som beslutas. Logiken bakom en ekonomisk utvärdering bygger på valmöjligheter, hur man ska välja och vilka konsekvenser dessa val ger.

En ekonomisk utvärdering görs med fördel i samband med en effektutvärdering men kan även genomföras efter att en effektutvärdering slutförts. Kapitlet behandlar inte ekonomiska utvärderingar som görs i tillägg till eller som efterföljande aktivitet när en effektutvärdering redan har planerats och genomförts. Här beskrivs snarare de åtgärder som behövs för att inkludera en ekonomisk utvärdering som en i förväg planerad studie med sitt eget egenvärde.

Vikten av ekonomiska utvärderingar

En ekonomisk utvärdering är ett viktigt verktyg för att förstå effektiviteten hos en intervention. Det beror på att den effektivaste interventionen vad gäller psykosociala effektutfall inte alltid är den mest kostnadseffektiva och ger därmed inte det bästa samlade utfallet för alla interventioner som genomförs med de resurser som finns till förfogande. Om det finns flera interventioner, till exempel med syfte att bekämpa olika problembeteenden bland ungdomar, som forskning visat är effektiva vad gäller psykosociala effektutfall – vilken av dem bör i så fall en kommun välja? Hur ska man hantera en situation där en viss intervention är dubbelt så effektiv att motverka vissa psykosociala problemtillstånd som alternativet, men fyra gånger dyrare att genomföra? Den intervention som är minst effektiv vad gäller behandling av enskilda individer kan vara den effektivaste när hela populationen behandlas, eftersom den kan vara den mest kostnadseffektiva. Att i ett sådant fall välja den mest effektiva avseende effektutfallet innebär en lägre effekt totalt sett och att resurserna inte används på bästa sätt. Den mest effektiva interventionen vad gäller psykosociala effektutfall är inte alltid den mest kostnadseffektiva. Men det är heller inte självklart att den billigaste interventionen är den mest kostnadseffektiva. För att kunna bedöma interventionernas effektivitet behövs data om både kostnader och effekter av interventioner.

Exempel på ekonomisk utvärdering inom det sociala området

På grund av höga kostnader och tveksamma behandlingsresultat har interventioner riktade till kriminella ungdomar skapat mycket uppmärksamhet i en kommun. Kommunen erbjuder tre interventioner för kriminella ungdomar.

Intervention X kostar 125 000 kr per person, intervention Y 80 000 och intervention Z 65 000. Kommunchefen vill lägga ner intervention X. Han anser att den är för dyr och att kommunen inte har råd att erbjuda en sådan dyr insats. Han tycker att alla kriminella ungdomar ska erbjudas intervention Z eftersom det är det billigaste alternativet.

Socialchefen påpekar att en effektutvärdering visar att med intervention X slutar 20 procent av ungdomarna med sin kriminalitet, med intervention Y 14 procent och med intervention Z fem procent. Socialchefen föreslår därför att de ska satsa på intervention X eftersom det är det mest effektiva alternativet och därför också det mest etiska valet oavsett kostnaden.

Chefsekonomen menar att intervention Y ska prioriteras. Den innebär att sju ungdomar måste behandlas för att en av dem ska sluta att vara kriminell. Det innebär en kostnad av 560 000 kr per lyckat fall (80 000 x 7). Motsvarande kostnad för intervention X är 625 000 kr per lyckat fall (125 000 x 5) och för Z 1 300 000 kr per lyckat fall (65 000 x 20).

Ett beslut fattas om att i första hand erbjuda intervention Y och intervention X till kriminella ungdomar, samt att upphöra att erbjuda intervention Z eftersom den visade sig vara inte bara den minst effektiva utan också den minst kostnadseffektiva.

Exemplet belyser risken med att basera besluten enbart på vilka interventioner som medför de lägsta kostnaderna eller ger de bästa resultaten. För att få en överblick över det relativa värdet av konkurrerande alternativ måste ekonomiska utvärderingar planeras och genomföras.

Design

Interventionsstrategier

Vid ekonomiska utvärderingar jämförs alltid två eller flera alternativa interventioner. Ofta jämförs en ny intervention med den intervention som normalt utförs (treatment as usual) eller med att ingen intervention erbjuds. Precis som i en effektutvärdering behöver de alternativ som ska studeras beskrivas så detaljerat som möjligt. Där-

utöver behövs information om relevanta kostnader (vem gör vad för vem, var och hur ofta samt till vilket pris) och effekter (vad är utfallet) för alla alternativ som ska studeras. I vissa fall får de interventioner som utvärderas olika resultat för olika delpopulationer. Det kan därför även vara relevant att beakta alternativa grupper av individer och presentera uppgifter om kostnader och effekter för respektive grupp. En studie som utvärderar en intervention för personer som är hemlösa kan till exempel också studera populationen utifrån missbruksproblem eller psykisk sjukdom. Det kan också vara relevant att studera interventionen utifrån olika intensitet, dos samt längd på interventionen.

Frågeställning

Alla ekonomiska utvärderingar har praktiska frågeställningar som måste besvaras. Vilken analysmodell som ska användas, vad som är en kostnad respektive effekt och hur dessa ska värderas beror på *vem* som ställer frågan och *varför*. En välgrundad ekonomisk analys inleds därför med att ställa en tydlig fråga så att den kan besvaras. En sådan frågeställning identifierar de alternativ som ska jämföras och vilka perspektiv som jämförelsen ska göras utifrån. Att fråga hur mycket en intervention kostar räcker inte för att göra en ekonomisk utvärdering. Frågeställningen måste behandla kostnader och effekter, valmöjligheter, alternativ och perspektiv. Exempel på tänkbara ekonomiska frågeställningar är:

- Ur kommunstyrelsens perspektiv, vilka är kostnaderna för och konsekvenserna av att placera barn som är utsatta för risk för övergrepp eller vanskötsel i fosterhem jämfört med att ge intensiv familjeterapi i hemmet?
- Ur socialtjänstens perspektiv, bör offer för upprepat våld i hemmet erbjudas intervention A eller intervention B?
- Ur ett samhällsperspektiv, vad är nettonyttan av att tillhandahålla föräldrautbildning till samtliga jämfört med att tillhandahålla föräldrautbildning enbart till högriskgrupper?

Det är viktigt att göra en operationell definition av det problem eller den fråga som ska analyseras. Den processen påverkar vilka typer av kostnader och effekter som ska inkluderas och bidrar till att fastställa vilken analysmodell som är lämplig att använda. En med omsorg och tydligt ställd fråga utgör grunden för övriga beståndsdelar i en ekonomisk analys. Beslut om analysens perspektiv, tidsram, analysmodellen, vilka kostnader som ska inkluderas och vilka konsekvenser som är av intresse fattas med bakgrund i den frågeställning som ska undersökas.

Perspektiv

Perspektivet för en ekonomisk utvärdering avgör vilka kostnader och effekter som ska tas med i undersökningen. Så långt som möjligt bör en ekonomisk utvärdering inom socialtjänstens område ha ett samhällsperspektiv, eftersom målet är att analysera hur samhällsresurserna mellan olika konkurrerande verksamheter kan användas på bästa sätt. För verksamheter som inte är skattefinansierade kan även andra perspektiv vara intressanta, till exempel ett verksamhetsperspektiv eller försäljningsperspektiv. I realiteten genomförs sällan ekonomiska utvärderingar med ett samhällsperspektiv, eftersom det sällan finns data eller resurser till detta. När perspektivet måste begränsas är det viktigt att begränsa det till det som är mest relevant för frågeställningen. Vid begränsning av perspektivet är en bra utgångspunkt att ange målgruppen för analysen: Identifiera vem som ska använda resultatet av analysen och ange hur resultatet ska användas. Utifrån detta kan den som gör undersökningen fastställa målgruppens informationsbehov avseende interventionen. Ett vidare perspektiv kan tillåta att utvärderingen även redovisas ur andra snävare perspektiv.

Samhällsperspektivet är det vidaste av perspektiven. I det analyseras alla kostnader och effekter för en intervention oavsett vem som betalar eller vem som berörs. Kostnaderna för en intervention ska spegla vad samhällsmedborgarna måste avstå från nu och i framtiden för att interventionen ska kunna implementeras. Det kallas

för alternativkostnad. Alternativkostnader inkluderar alla finansiella och icke-finansiella kostnader för en intervention, oavsett om de uppstår hos en viss aktör, verksamhet eller individ. Uppgifter om kostnaderna kan därför behöva samlas in från flera håll, till exempel från skolan, individer och socialtjänst, för att få fram en fullständig bild av kostnaderna.

Alternativkostnaderna inkluderar mer än bara de finansiella eller budgetmässiga utlägg som hör samman med leverans eller mottagande av tjänster. De innefattar även de resurser som inte blir tillgängliga för samhället på grund av uteblivet arbete eller förlorad ”frisk” tid. Det kallas för produktionsbortfall. Produktionsbortfall är inte en reell kostnad och syns inte i någon budget. Denna alternativkostnad behöver beräknas genom exempelvis humankapital- eller friktionsmetoden. Interventioner som inbegriper uppsökande verksamhet i samhället kan innefatta volontärer eller inkludera lönebidrag till viss personal från andra källor. Implicita kostnader eller skuggpriser måste tas fram för att värdera volontärers tid och budgeten kan behöva justeras för att spegla de externa bidragen så att dessa kostnader bedöms korrekt utifrån samhällsperspektivet. En viktig fördel med samhällsperspektivet är att man undviker risken att identifiera alternativ som kostnadseffektiva om de enbart innebär en förskjutning av kostnader mellan sektorer eller grupper. Studeras enbart en verksamhets effektivitet finns risken att samhällets effektivitet försämras. Det kan ske om till exempel kostnaderna för en intervention finns i en verksamhet medan effekterna av att inte utföra interventionen kostar för någon annan verksamhet.

> ***Perspektivet förändrar kostnadsuppskattningen***
> I en ekonomisk utvärdering genererar alla resursobjekt (personal, lokaler, bilar, material osv.) för en intervention en kostnad, oavsett om en viss aktör verkligen har betalat för den eller inte. I ekonomiska analyser gäller med andra ord att ingenting är gratis. Beroende på vilket perspektiv som tillämpas kan kostnaden vara dold och måste då uppskattas. Ta till exempel en volontär. En sådan person ger ett verkligt bidrag till hur interventionen fungerar och är därför en värdefull resurs för organisationen, oavsett om

personen får lön eller inte. Det teoretiskt rätta sättet att värdera personens insats är att estimera vad individens tid är värd om han eller hon skulle sysselsätta sig med den näst bästa alternativa aktiviteten. På motsvarande sätt kan en jourtelefonfunktion användas för att identifiera personer som behöver hjälp och som refereras vidare till en intervention, men den kanske finansieras genom donationer eller andra filantropiska källor. Värdet av dessa bidrag ska räknas till den totala kostnaden för interventionen.

Ett annat exempel är ett samarbete mellan det kommunala socialkontoret och skolan där vissa av socialtjänstens klienter går i en specialklass. Socialkontorets kostnad för interventionen rör endast tillhandahållandet av lokalen medan skolan står för lärarkostnaden. I en ekonomisk utvärdering är det viktigt att ta hänsyn till både socialkontorets och skolans kostnader.

En liknande situation är den där kommunstyrelsen finansierar ett bostadsprogram för personer med alkohol- eller drogberoende. Det lokala socialkontoret kan hänvisa sina klienter till det bostadsprogrammet utan kostnad för socialkontoret. I en ekonomisk utvärdering ska samtliga kostnader redovisas, oavsett vem som betalar notan.

Interventioner som överför pengar från en grupp människor till en annan, till exempel från den del av befolkningen som arbetar till dem som inte har arbete och behöver försörjningsstöd, förändrar inte det sammanlagda värdet av de resurser som är tillgängliga för samhället. Ingen alternativkostnad uppstår. Denna omfördelning av pengar kallas transferering (eng. *transfers*), och ska normalt inte tas med i en ekonomisk utvärdering som görs ur ett samhällsperspektiv eftersom de inte utgör några reella kostnader för samhället. De kan emellertid vara viktiga ur beslutsfattarens perspektiv, då de får verkliga konsekvenser för socialkontorets budget. Även om utbytet av pengar i sig inte nödvändigtvis innebär att resurserna har förbrukats, innebär det att färre resurser är tillgängliga för det betalande socialkontoret. Trots det bör en ekonomisk analys följa upp och redovisa överföringar om de är betydande, eftersom omfördelningseffekterna av en intervention ofta är av intresse för målgruppen. När överföringskostnaderna beskrivs är det viktigt att betona att de inte bör läggas till de reella kostnaderna för samhällsresurserna i analysen, utan de ska redovisas vid sidan om.

Att begränsa perspektivet för kostnaderna är det absolut vanligaste tillvägagångssättet i de rapporter som publicerats om kostnadsdata i randomiserat kontrollerade utvärderingar (Barber & Thompson, 1998). Brouwer, van Exel, Baltussen och Rutten (2006) föreslår ett

tillvägagångssätt med två perspektiv i alla rutinmässiga analyser, där ett resultat bygger på leverantörens perspektiv och ett annat på det vidare samhällsperspektivet. Drummond, Sculpher, Torrance, O'Brien och Stoddare (2005) föreslår att analyser i tveksamma fall ska göras ur samhällsperspektivet, som är det vidaste och det som alltid är relevant.

Tidsramar

När en ekonomisk utvärdering planeras i samband med en effektutvärdering är det viktigt att planera tids- och analysramar. Tidsramen för en utvärdering är den period under vilken interventionerna genomförs och analysramen är den period under vilken de kostnader och effekter som uppstår till följd av interventionen beaktas. Analysramen sträcker sig därmed ofta längre än tidsramen, eftersom effekterna av en intervention kan fortgå efter det att interventionen har avslutats. Det kan till exempel gå många år mellan preventiva interventioner och effekter i form av uteblivna sociala problem. Effekterna av prevention kan med andra ord inträffa långt fram i tiden.

När primärdata ska samlas in är det viktigt att noga överväga tidsaspekten för den ekonomiska utvärderingen. Effektutvärderingar använder sig oftast av tidpunktsbaserade datainsamlingsmodeller. Då samlas data in i samband med till exempel intagningen, efter sex månader, efter två år och efter fem år. Ekonomiska utvärderingar behöver dock kontinuerliga datainsamlingsmodeller som pågår under hela undersökningens tidsram. Man måste till exempel dagligen notera om undersökningspersonerna har deltagit i olika aktiviteter för att kunna uppskatta kostnaden för interventionen och för att kunna värdera effekterna i en prospektiv analys. Det är viktigt att komma ihåg att den valda tidsramen bör:

1. vara samma för de kostnader som samlas in som för dem som deltar i interventionen
2. vara tillräckligt lång för att undvika eventuella sekulära mönster (t.ex. sommarsemester som kan påverka antalet klienter som har

fått intervention samt deltagandet i interventioner och utvärderingen)
3. vara tillräckligt lång för att fånga upp programmets start- och underhållskostnader.

Även om startkostnaderna (t.ex. utbildning av personal) kan vara viktiga att redovisa av implementeringsskäl är det inte alltid så att dessa kostnader ska ingå i den uppskattade enhetskostnaden. Det beror på att startkostnaderna kan vara försumbara när de sprids ut över interventionens livslängd och om de tas med skulle det i så fall leda till att uppskattningen av marginalkostnaden blir överdrivet hög.

Prospektiv, retrospektiv och hybridmetod för datainsamling

En prospektiv ekonomisk utvärdering följer upp resursanvändningen och kostnaderna när de uppstår och säkerställer att ekonomiska data finns tillgängliga vid den tidpunkt när försöket avslutas. Med den här metoden kan uppgifter om direkta, indirekta, produktivitetsrelaterade och immateriella kostnader samlas in. En prospektiv insamling av ekonomiska data kräver dock omfattande planering och skiljer sig i det avseendet inte från prospektiv insamling av kliniska data i en effektutvärdering.

Vid en retrospektiv ekonomisk utvärdering uppskattas resursanvändningen och kostnaderna efter att de har uppstått och det innebär ett mer begränsat tillvägagångssätt för att bedöma resursanvändningen där man normalt koncentrerar sig på direkta kostnader eftersom det ofta saknas tillgång till övriga typer av data.

Valet mellan prospektiv och retrospektiv datainsamling vilar normalt på den förväntade psykosociala effekten av en intervention och på hur stora resurser utredarna har. Om möjligt är ett prospektivt tillvägagångssätt att föredra eftersom det möjliggör en mer ingående analys av kostnaderna för en intervention. I realiteten är många ekonomiska utvärderingar en hybrid där data om interventionen av intresse exempelvis samlas in prospektivt under undersökningens gång och övriga uppgifter om kostnader och effekter samlas in retrospektivt. På motsvarande sätt kan resursanvändningen samlas in prospektivt i samband med utvärderingen och värdet på resurserna beräknas retrospektivt.

Analysmodell

Val av analysmodell beror på frågeställningen, vilka resultat som är av intresse och vilken information som är tillgänglig. Det finns tre vanliga metoder för ekonomiska utvärderingar av sociala interventioner:

1. Kostnads- och intäktsanalys (eng. *cost-benefit analysis*) är en teknik där både kostnaderna och effekterna uttrycks i monetära termer. Nettonuvärde används ofta för att sammanfatta resultatet.
2. Kostnads- och effektivitetsanalys (eng. *cost-effectiveness analysis*) är en teknik där kostnaderna uttrycks i monetära termer och effekten i effektenheter. För denna metod krävs att endast en effekt studeras eller att effekterna kan vägas samman. Mått som kostnadseffektkvot eller kostnad per effektenhet används ofta för att sammanfatta resultatet.
3. Kostnads- och effektanalys (eng. *cost-consequence analysis*) påminner om kostnads- och effektivitetsanalys. I den redovisas kostnader i monetära termer och effekterna i effektenheter separat. Inget försök görs att summera resultatet.

Dessa tre metoder hanterar kostnaderna på samma sätt, men analyserar och redovisar effekterna på olika sätt. Valet av lämplig analysmodell beror på målgruppen för studien, frågeställningen och tillgängligheten till data. I vissa fall kan fler än en metod användas i samma studie för att besvara specifika frågeställningar. I de flesta fall är kostnads- och effektanalysen det lämpligaste valet när det gäller social omsorg. Det beror på att sociala interventioner vanligtvis är utformade för att påverka en rad olika resultat som mäts med olika skalor, vilket inte gör det så lätt att omvandla dessa till ett index eller ett monetärt värde. Inom hälsoekonomin, som studerar hälso- och sjukvården, används ofta metoden kostnads-nyttoanalys (eng. *cost-utility analysis*) som motsvarar kostnads- och effektivitetsanalysen. Då mäts effekterna i till exempel kvalitetsjusterade levnadsår (eng. *Quality Adjusted Life Year, QALY*) eller funktionsjusterade levnadsår (eng. *Disability Adjusted Life Year, DALY*), som inte till fullo

är användbara inom socialvården. Kostnads- och intäktsanalysen skulle kunna vara ett bra alternativ, men den är jämförelsevis dyr och det är ofta svårt att omvandla effekterna till monetära värden. Vad skulle samhället till exempel betala för en förändring med en procent i Child Behavior Checklist (Achenbach, 1991)?

Det finns ett teoretiskt samband mellan val av perspektiv och val av analysmodell för en ekonomisk utvärdering. Vid kostnads- och intäktsanalys tillämpas i allmänhet ett samhällsperspektiv. I praktiken används dock kostnads- och intäktsanalyser sällan inom den sociala omsorgsverksamheten (Sefton, Byford, McDaid, Hills & Knapp, 2002). Perspektivfrågan är något mer komplicerad när det gäller kostnads- och effektivitetsanalyser och kostnads- och effektanalyser, eftersom kostnader och konsekvenser inte mäts i samma enhet.

Statistisk styrka

Antalet undersökningspersoner i en effektutvärdering ska baseras på powerberäkning för att kunna dra säkra slutsatser om eventuella effekter. Antalet undersökningspersoner som krävs för den ekonomiska utvärderingen kan dock vara större än vad som krävs för effektutvärderingen. Resursanvändningen och kostnaderna tenderar ofta att vara snedfördelade, vilket innebär en större varians för kostnadsvariablerna än för de kliniska effekterna. Vid jämförelser av olika interventioners kostnader krävs därför vanligtvis ett större urval än för motsvarande jämförelser av psykosociala effekter. Även om själva effektutvärderingen har hög power kan med andra ord den ekonomiska utvärderingen ha låg power om den inte planeras i samband med effektutvärderingen. Problemet är i de flesta fall en lägre precision i kostnadseffektiviteten, eftersom det måttet påverkas av både täljaren och nämnaren, och en inexakt nämnare ökar variationen i måttet (Gafni, Walter, Birch & Sendi, 2008; Walter, Garnmi & Birch, 2007).

Andra faktorer att beakta när det gäller urvalets storlek är hur stor skillnad i kostnader som är meningsfull ur ett ekonomiskt per-

spektiv, hur kostnaderna fördelar sig och tillförlitligheten i de mätmetoder som används för att analysera kostnaderna.

Mätinstrument

Värdering av utfall

Alla ekonomiska utvärderingar redovisar kostnader i ett penningvärde. Analyserna behandlar däremot de psykosociala effekterna olika. I en undersökning där en kostnads- och intäktsanalys tillämpas uttrycks till exempel alla psykosociala effekter i monetära termer medan i en kostnads- och effektivitetsanalys uttrycks de i en effektenhet och i en kostnads- och effektanalys i flera effektenheter. I samband med planeringen av en effektutvärdering är det därför viktigt att överväga vilka instrument som ska användas för att mäta de psykosociala effekterna för den ekonomiska utvärderingen. Om målet med den ekonomiska utvärderingen är att värdera effekter monetärt måste man välja resultatmått som det går att beräkna penningvärdet av. Mått för uppföljning av förändringar i kriminalitet kan till exempel vara lättare att prissätta än mått som beskriver ett mer allmänt antisocialt beteende. Om målet är att redovisa en kostnad per effektenhet gäller på motsvarande sätt att noggrant överväga vilket resultatmått som ska väljas som effektmått eller hur flera effekter kan vägas samman till ett mått. Eftersom effekterna kan vara såväl positiva som negativa, kan uteslutandet av vissa resultatvariabler medföra att man missar viktiga negativa konsekvenser och den ekonomiska utvärderingen som redovisas blir då ofullständig, felaktig och därmed missvisande.

Metod för värdering av utfall

Trovärdigheten i en effektutvärdering beror på studiens interna validitet. Den interna validiteten av effektutvärderingen får därmed en direkt effekt på resultatet av den ekonomiska utvärderingen. Med låg intern validitet i effektutvärderingen ökar osäkerheten i den ekonomiska utvärderingen. Om en effektutvärdering planeras ha lägre

intern validitet bör detta påverka hur resultatet bedöms i den ekonomiska utvärderingen. För utvärderingar som exempelvis har ett kvasiexperimentellt upplägg, i stället för ett experimentellt, kan sensitivitetsanalys användas för att testa hur olika effektstyrkor skulle påverka det ekonomiska resultatet.

Datainsamling

Generellt finns det bara två typer av uppgifter som behövs för att kunna göra en ekonomisk utvärdering i samband med en effektutvärdering: information om resursanvändning och enhetspris.

Resursanvändning

Inom ramen för effektutvärderingar sker resursanvändningen när deltagarna (a) involveras i respektive intervention som utvärderas och (b) använder andra efterföljande resurser under studiens tidsram. Studien måste därför följa upp resurser i form av exempelvis behandlingstid, restid och tid som deltagaren ägnar åt intervention. Beroende på tidsramen, studiens perspektiv, frågeställningen och den typ av analysmodell som används kan resursanvändningen behöva följas upp för enskilda deltagare med avseende på:

1. den intervention som utvärderas (t.ex. den intervention som behandlingsgruppen får)
2. förutsättningarna för jämförelsen (t.ex. den intervention som kontrollgruppen får)
3. övriga tjänster (t.ex. andra interventioner som gavs individer under tidsramen för att komma till rätta med problemet).

Nedan följer exempel på datainsamlingsmetoder för att beräkna deltagarnas resursanvändning:

- Loggböcker kan användas som en prospektiv datainsamlingsmetod. Den verksamhetsansvarige kan använda loggböcker för att följa upp deltagarnas inblandning i interventionen. En verksamhetsansvarig kan till exempel notera datum samt start- och slut-

tid för enskilda behandlingar eller gruppaktiviteter för att mäta det totala antalet behandlingstimmar eller -minuter och vem som utför behandlingen.
- Dagböcker är en annan prospektiv datainsamlingsmetod som påminner om loggböcker, men fylls i av deltagarna själva. De kan användas för att följa upp olika resursanvändning, såsom frånvaro från arbetet, resekostnader och deltagandet i behandlingar.
- Aktdata kan användas som en retrospektiv datainsamlingsmetod. I vissa fall kan akter vara lika korrekta som loggböcker. Det förekommer dock att informationen inte registreras på lika detaljerad nivå som i dagböcker eller loggböcker.
- Administrativa journaler eller databaser kan också användas som en retrospektiv datainsamlingsmetod. Administrativa journaler och databaser kan förekomma både i aggregerad och icke-aggregerad form.

När informationen om deltagarnas resursanvändning har samlats in måste varje resurskategori tilldelas ett värde. Om målet är att göra en kostnads- och intäktsanalys måste man även försöka värdera de utfallsområden som ingår i effektutvärderingen. De mätinstrument som väljs för effektutvärderingen kan också användas som instrument för att uppskatta förändringar i resursanvändning. En effektutvärdering kan till exempel följa förändringar i kriminalitet, drogmissbruk eller hemlöshet. Genom dessa förändringar kan en uppskattning av förändringen i resursanvändning göras, eftersom en förändring av exempelvis kriminalitet får ekonomiska konsekvenser för till exempel enskilda individer, brottsoffren, produktiviteten samt polis- och domstolsväsende. En ekonomisk utvärdering med ett samhällsperspektiv måste inkludera alla de skillnader i psykosociala utfall som kan identifieras. En kostnads- och intäktsanalys bör dessutom översätta dessa förändringar till ett penningvärde när de inträffar.

Enhetspris

Det kan ibland vara komplicerat att fastställa enhetspriset för de resurser som följs i den ekonomiska utvärderingen, eftersom det ofta inte finns någon självklar metod för prissättning. Vad kostar en timme av deltagarnas tid för att söka upp och genomgå interventionen eller tillhandahållande av en stödgruppsintervention till exempel? Idén om att uppskatta enhetspriser kan tyckas komplicerad, eftersom det innebär att data samlas in på organisationsnivå och därmed ligger utanför ramarna för det urval av personer som används för att avläsa interventionens psykosociala effekter och resursanvändningen.

Ett första steg vid uppskattning av enhetspriset för en resurs är att veta vilka specifika resursobjekt som behövs för att implementera interventionerna. Ett användbart tillvägagångssätt är att dela in interventionerna i funktionella kategorier, till exempel

- personal
- lokaler
- utrustning
- material
- transport
- utbildning
- publicitet, information och kommunikation.

Andra kategorier kan förstås också användas. Inom exempelvis socialtjänsten är personalresurser den största kostnaden för interventioner. Man kan därför välja att bryta ner kostnaderna i två kategorier, personalkostnader och övriga kostnader. Det fungerar dock inte för interventioner som har andra stora kostnadskategorier, till exempel hälso- och sjukvården.

När man har identifierat de resursobjekt som behövs för att tillhandahålla en intervention är nästa steg att:

1. mäta hur mycket av varje resursobjekt som används för att tillhandahålla interventionen under en viss tidsperiod (t.ex. ett år eller en termin)

Kostnader

Den befintliga litteraturen om ekonomiska utvärderingar inom hälso- och sjukvården innehåller flera olika definitioner av kostnader. Kostnader kan till exempel vara direkta, indirekta eller immateriella, de kan vara finansiella eller ekonomiska, de kan vara engångskostnader eller löpande, de kan vara fasta eller rörliga. Kostnader kan dessutom vara totala, årliga, genomsnittliga eller marginalkostnader. Dessa grupperingar utesluter inte varandra (Wonderling, Gruen & Black, 2007) och tillämpningen av dem kan ibland vara inkonsekvent (Drummond m.fl., 2005). Att förstå hur olika kostnader skiljer sig från varandra och hur de olika kostnadskategorierna kan användas för att klassificera kostnader i samband med en ekonomisk utvärdering underlättar för utredaren och garanterar att alla kostnader tas med i analysen. Att dessutom klart och tydligt förstå målet med en viss ekonomisk utvärdering kan underlätta beslutet om vilka kostnader som ska inkluderas och beaktas i analysen.

2. sätta ett värde på varje resursobjekt så att den totala kostnaden för interventionen under den valda tidsperioden kan fastställas
3. mäta hur många enheter (behandlingar, dagar, timmar, minuter osv.) som kan tillhandahållas under tidsperioden till den fastställda totala kostnaden (2 ovan).

Därefter kan ett enhetspris beräknas för varje uppmätt resurs genom att dividera den totala kostnaden (2 ovan) med antalet tillhandahållna enheter (3 ovan).

$$\textit{Enhetspris} = \frac{\text{Den totala kostnaden under perioden}}{\text{Antal enheter producerade under perioden}}$$

Det enklaste sättet att värdera resurserna är att använda penningvärdet som mått även om det mest teoretiskt korrekta måttet är alternativkostnaden. I många fall har den som levererar interventionen data om de enskilda resursobjekt som behövs för interventionen i bokföringssystemet. I andra situationer är det inte lika lätt att identifiera resursobjekten och uppskatta värdet av dem.

Några av de metoder som kan användas för att identifiera, samla in och värdera resurser för att uppskatta enhetspriset beskrivs kortfattat nedan.

Direkt skattning

När det inte går att identifiera en enskild individ eller intervention genom befintliga datakällor eller dessa datakällor är otillräckliga på annat sätt, kan data samlas in genom enkäter och observationer. Det kallas för direkt mätning. Vanliga metoder för direkt mätning är följande:

- *Tidsstudier* (eng. *time in motion studies*). Utbildade tidsstudiemän observerar personal och klienter för att fastställa hur mycket tid som ägnas åt interventionen.
- *Loggböcker*. Personalen fyller i en logg med aktiviteter som är kopplade till interventionen.
- *Dagböcker*. Klienterna fyller i en enkät om hur mycket tid som ägnas åt direkt vård, transporter och andra resurser relaterade till interventionen.

Indirekta mått

I vissa fall finns det indirekta mått (eng. *proxy measures*) att tillgå. I dessa fall måste man söka andra källor än de som tillhandahåller interventionen. Kanske finns liknande interventioner som tillhandahålls av andra organisationer eller en extern finansiär som har information om priset som betalas för en intervention. På liknande sätt kan det finnas data inom hälso- och sjukvården, psykiatrin eller andra förvaltningar och myndigheter för liknande interventioner eller åtgärder som kan användas som en uppskattning av enhetspriset.

Arbetsledarenkäter

Arbetsledarna fyller i en enkät som exempelvis är till för att uppskatta tiden som läggs ner på interventionen av respektive personalgrupp, rådgivare, socialarbetare, sjuksköterskor och så vidare.

Statistiska kostnadsmodeller

För att kunna uppskatta kostnaderna med hjälp av en statistisk kostnadsfunktion måste utredaren ha en datakälla som inkluderar kostnaderna och de faktorer som till största del kan förklara variationen i kostnaderna. Enhetspriserna kan sedan uppskattas med hjälp av grundläggande ekonometriska metoder.

Oavsett vilken metod som väljs är det viktigt att överväga hur metoden påverkar uppskattningen av enhetspriset. Blir uppskattningen för hög eller låg? Det kan testas med känslighetsanalys.

Även i sammanhang där ekonomiska analyser och kostnadsbedömningar prioriteras tillämpar de flesta utredarna en hybridmodell (Barnett, 2003), där en metod används för en del av undersökningen och en annan metod för övriga delar av undersökningen. Om man exempelvis använder en mer exakt metod för att bedöma enhetspriser för kostnader och en mindre exakt metod för att bedöma enhetspriser för effekterna kan det leda till en analys som inte bara är konservativ när det gäller forskningens resursanvändning utan även när det gäller det redovisade resultatet (Olsson, 2011).

Analys

Aggregering

När data om resursanvändning har samlats in för varje deltagare och enhetspriset har uppskattats för varje resurs som har följts upp i studien kan de totala kostnaderna och effekterna beräknas för varje deltagare och ett genomsnitt räknas fram för att få fram den genomsnittliga kostnaden för behandlingsgruppen och jämförelsegruppen. För en enskild deltagare i studien beräknas kostnaden för en intervention genom att multiplicera mängden använda resurser med enhetspriset.

$$\textit{Total kostnad} = \textit{Resursanvändning} \times \textit{Enhetspris}$$

I de situationer där deltagarna använder en rad resurser eller där effekterna ska värderas i pengar måste resursanvändningen och enhetspriset uppskattas separat för varje resurs:

$$\textit{Total kostnad} = (Q_1 \times P_1) + (Q_2 \times P_2) + \cdots + (Q_n \times P_n)$$

där
Q_1 = använd mängd av resurs 1
P_1 = enhetspris för resurs 1

Aggregeringen av kostnaderna för alla deltagare görs med hjälp av följande grundläggande ekvation:

$$TC = \sum_{i=1}^{N} (Q_1 \times P_1) + (Q_2 \times P_2) + \cdots + (Q_n \times P_n)$$

där
TC = den totala kostnaden för deltagare i
Q_1 = använd mängd av resurs 1
P_1 = priset för resurs 1
i = deltagare

Statistisk analys

Eftersom ekonomiska utvärderingar handlar om att förstå den ekonomiska effekten av konkurrerande interventioner är det viktigt att särskilja faktiska skillnader från dem som enbart beror på tillfälligheter. Därför är hypotestest lika viktigt för den ekonomiska utvärderingen som för effektutvärderingen. För de ekonomiska utvärderingarna är det emellertid viktigt att tänka på att många antaganden som är nödvändiga för parametrisk hypotestestning inte gäller. Till exempel kännetecknas kostnadsdata av att de är snedvridna. Det beror på att kostnadsdata har en naturlig nollpunkt men ingen naturlig övre gräns. Dessutom förekommer ofta några få fall som har relativt höga totalkostnader. Det finns alternativ för att hantera data som bryter mot antagandena för parametriska test, till exempel att

använda ett summerande mått i stället för medelvärdet som medianen och testa placeringen av det måttet, eller att använda icke-parametriska test. Det finns dock kritik mot dessa alternativ (Barber & Thomson, 1998; Briggs & Gray, 1999; O'Hagan & Stevens, 2003). Det är därför viktigt att redovisa så många detaljer om kostnadsfördelningen som möjligt utifrån de osäkerheter som omger hypotestestmetoden för ekonomiska data.

Vad gäller hypotestestning bör utredaren ha i åtanke att den osäkerhet som man är villig att acceptera när det gäller det kliniska resultatet kanske inte är samma som den osäkerhet man är villig att acceptera i den ekonomiska utvärderingen. När en intervention har befunnits vara signifikant effektivare än alternativet att bekämpa vissa psykosociala problem kan en beslutsfattare vara mycket mer villig att godta ett mer generöst p-värde än 0,05 till exempel, för bedömningen av det ekonomiska resultatet. Även om jämförelser av medelvärden ger mest information inför ett beslut bör konfidensintervall redovisas som kompletterande information för att bättre kunna bedöma den risk som ett visst ekonomiskt resultat medför.

Aggregering av kostnader för samtliga deltagare

I det här hypotetiska exemplet undersöks de inkrementella kostnaderna (förändring i kostnaderna) och effekterna för tre interventioner riktade till ungdomar med missbruksproblem. Den första interventionen A är en intensiv hemmabaserad intervention. Intervention B är en rådgivningsintervention med regelbunden drogtestning. Den sista interventionen C är en placeringsintervention med rådgivning.

De två första deltagarna i respektive interventionsgrupp förbrukade följande resurser under den tre månader långa utvärderingsperioden: Deltagare D1 fick intervention A under 120 timmar. D2 fick 48 timmar av intervention A. D3 och D4 fick intervention B; D3 deltog i nio sessioner à en timma och tre drogtester, medan D4 deltog i fem sessioner à en timma och fem drogtester. D5 och D6 ingick i intervention C under tremånadersperioden och deltog varje vecka i rådgivningssessioner.

En uppskattning av enhetspriset för varje resurs har gett följande resultat:

Resurs	A	B	C
Placering	0 kr	0 kr	2 000 kr/dag
Rådgivning	1 000 kr/tim	500 kr/tim	300 kr/tim
Testning	0 kr	300 kr/test	0 kr
Deltagarnas resor	0 kr	100 kr/resa	0 kr

Aggregeringen av kostnader per deltagare i studien ser ut så här:

	Resurs 1	Resurs 2	Resurs 3	Total kostnad
D1	120 tim*1 000 kr	0	0	120 000 kr
D2	48 tim*1 000 kr	0	0	48 000 kr
D3	9 tim*500 kr	18 resor*100 kr	3 st*300 kr	72 000 kr
D4	5 tim*500 kr	10 resor*100 kr	5 st*300 kr	5 000 kr
D5	90 tim*2 000 kr	12 tim*300 kr	0	183 600 kr
D6	90 tim*2 000 kr	12 tim*300 kr	0	183 600 kr

Den genomsnittliga kostnaden per deltagare i respektive interventionsgrupp kan beräknas genom att summera de totala kostnaderna för varje deltagare i respektive grupp och dividera dem med antalet deltagare (2 i det här exemplet):

Intervention A	120 000 + 48 000/2 = 84 000 per deltagare
Intervention B	72 000 + 5 000/2 = 38 500 per deltagare
Intervention C	183 600 + 183 600/2 = 183 600 per deltagare

En grundlig analys skulle visa vem som betalar vad, till exempel intervention B 1 200 kr i genomsnitt är kostnader som den enskilda deltagaren själv får stå för och 26 500 kr i genomsnitt är kostnader som kommunen står för.

Tidpunkt för kostnader och konsekvenser

Beslutet att implementera en viss intervention medför konsekvenser som sträcker sig framåt i tiden. Värdet av kostnaderna och konsekvenserna för dessa beslut varierar från år till år, eller från en tidsperiod till en annan. Justeringen av dessa värden är särskilt intressant för den ekonomiska utvärderingen av preventiva interventioner där kostnaderna vanligtvis uppstår tidigt under försöket och effekterna dyker upp först långt fram i tiden. Om en ekonomisk utvärdering

sträcker sig över mer än ett år måste två justeringar – dels för inflation, dels för tidpreferensen – göras av värdet av de kostnader och effekter som inte uppstår under basåret.

Inflation

När data som avser mer än en tidsperiod, vanligtvis ett år, samlas in måste de inflationsjusteras innan de analyseras. Det beror på att valutans köpkraft förändras med tiden. Den valuta som data samlas in i kallas nominell. Värdet på, eller köpkraften för, den nominella valutan förändras från år till år. Köpkraften för 100 kronor år 2010 är till exempel mycket mindre än den var 1970. För att kunna räkna med den avtagande köpkraften hos en valuta på grund av inflationen måste de nominella värdena räknas om till reella värden (deflateras).

Det finns flera prisdeflatorer och vilken prisdeflator man väljer beror på detaljerna i den undersökning som görs. Ett exempel är om handel bedrivs på den öppna marknaden med de tjänster som studeras. Om så är fallet kan konsumentprisindex (KPI) eller bruttonationalprodukt (BNP) vara vettiga alternativ. Dessa deflatorer är baserade på konsumentvaror och tjänster och är avsedda att mäta förändringar av levnadskostnaderna från en period till en annan. Om tjänsterna däremot produceras av en myndighet och tillhandahålls utan kostnad kan producentprisindex (PPI) vara ett bättre alternativ. Till skillnad från KPI och BNP mäter PPI förändringar av kostnaden för att producera varor och tjänster i stället för kostnaden för att köpa dem. Utöver dessa index finns för kommunernas och landstingens verksamheter konsumtionsindex och produktionsindex som Statistiska centralbyrån tar fram årligen.

Omvandlingen av värden från nominella till reella sker med hjälp av följande ekvation:

$$RV_b = \frac{NV_a}{PD_a} \times PD_b$$

där

RV_b = valutans reella värde år b
NV_a = valutans nominella värde år a
PD_a = prisdeflatorns värde år a
PD_b = prisdeflatorns värde år b

En valutas värde kan antingen skrivas tillbaka, deflateras, eller skrivas fram, inflateras, med hjälp av en prisdeflator. Det som är viktigt här är inte i vilken riktning justeringen görs, utan att alla värden redovisas (år och vilken valuta) för samma basår eftersom det gör det möjligt att överföra resultatet från en situation till en annan.

Inflatering av priser till basåret

År 2009 gjordes en studie för att analysera kostnaden för sociala interventioner under en tvåårsperiod (Olsson, 2010). Deltagandet i studien pågick under en ettårsperiod och undersökningsperioden sträckte sig därför över fyra kalenderår, 2004–2007. De ekonomiska variabler som värderades i nominell valuta mellan 2004 och 2007 inflaterades till reella värden 2007 med hjälp av producentprisindex (Sveriges officiella statistik, 2008). Alla (avrundade) värden redovisas i svenska kronor.

År	PPI	Nominellt SEK	Reellt SEK
2004	123,5	40 000	45 300
2005	128,0	231 700	252 900
2006	133,6	414 900	433 800
2007	139,7	10 300	10 300
Totalt		*696 900*	*742 300*

Tidspreferens

Diskontering utförs eftersom individer fäster större värde vid att ha saker, till exempel varor, tjänster, pengar och hälsa, i dag än i framtiden. För att kunna spegla den tidspreferensen diskonteras framtida kostnader för och effekter av en intervention tillbaka till det datum när en individ togs med i en undersökning. På så sätt speglas värdet av framtida kostnader och effekter vid intagningen. Den diskonteringsränta som används ligger vanligtvis på mellan tre och

fem procent. I vissa länder föreskriver regeringen att en viss diskonteringsränta används när ekonomiska utvärderingar görs av projekt som finansieras med offentliga medel.

Diskonteringen underlättas av formeln nedan som används för att beräkna nuvärdet (eng. *present value*):

$$NV = \sum_{t}^{N} FV \times \frac{1}{(1 + r)^t}$$

där

NV = nuvärde
FV = framtida värde
r = diskonteringsränta
t = tidsperiod

Den diskonteringsränta som väljs påverkar resultatet. Av den anledningen vållar diskontering debatt mellan olika ekonomer. Det är därför lämpligt att redovisa både ett icke-diskonterat och ett diskonterat resultat och att testa diskonteringsräntan med hjälp av en känslighetsanalys.

Diskontering av framtida kostnader och effekter (jfr Morris, Devlin & Parkin, 2007)

En intervention kostar 100 000 kronor per år att införa under fem år (efter justering för inflation). Dessa kostnader måste läggas ihop för att kunna beräkna den totala kostnaden som sedan vägs mot effekterna av interventionen. Eftersom kostnaderna uppstår under olika tidsperioder måste de årliga kostnaderna diskonteras för att omvandla dem till motsvarande nuvärde.

Med "t_0" avses kostnader som uppstår den första ettårsperioden (eftersom diskonteringen vanligtvis görs årsvis). Med t_1 avses det andra året och så vidare. Om diskonteringsräntan är 5 procent är $r = 0{,}05$. När värdena på r och t stoppas in ser ekvationen ut så här:

t_0	t_1	t_2	t_3	t_4
100 000 x 1/	100 000 x 1/	100 000 x 1/	100 000 x 1/	100 000 x 1/
$(1 + 0{,}05)^0$	$(1 + 0{,}05)^1$	$(1 + 0{,}05)^2$	$(1 + 0{,}05)^3$	$(1 + 0{,}05)^4$

En beräkning av vart och ett av de fem åren ger följande (avrundade) nuvärden:

t_0	t_1	t_2	t_3	t_4
100 000	95 200	90 700	86 400	82 300

När dessa belopp summeras får man ett nuvärde på 454 600 kronor att jämföra med de 500 000 kronor som är resultatet utan diskontering.

Presentation av det ekonomiska resultatet

Den ekonomiska utvärderingen sammanfattar skillnaderna mellan kostnaderna för de olika interventioner som jämförts och skillnaderna mellan effekterna för de olika alternativen. I en kostnads- och effektanalys redovisas till exempel kostnad och effekter per deltagare utan något försök att sammanfatta vidare. Med en kostnads- och effektivitetsanalys redovisas effekten i effektenheter utan något försök att prissätta den tillsammans med kostnaden. Kostnadseffektkvot eller kostnad per effektenhet brukar användas för att sammanfatta resultaten. I en kostnads- och intäktsanalys kan det ekonomiska resultatet sammanfattas genom nettonuvärdet.

Nettonuvärde

Nettonuvärdet (eng. *net present value*) för en intervention utgör skillnaden mellan nuvärdet av interventionens effekter och nuvärdet av kostnaderna för interventionen. Beräkningen av nettonuvärdet görs med ekvationen:

$$\mathrm{NNV} = \sum_{t=0}^{N} \frac{\mathrm{N}_t}{(1-i)^t} - \sum_{t=0}^{N} \frac{\mathrm{C}_t}{(1-i)^t}$$

där

NNV = diskonterat nuvärde för interventionen
N_t = effekterna (intäkten, nyttan) som uppstår under tidsperioden t
C_t = kostnaderna som uppstår under tidsperioden t
i = räntan

I vissa texter föreslås att intäktskostnadskvoten eller internräntan för avkastningen används som en lämplig summering av det ekonomiska resultatet av en kostnads- och intäktsanalys. Boardman med flera (2006) visar hur dessa summeringsmått kan ge ett missvisande resultat, särskilt om de åtgärder som jämförs har olika omfattning. De menar att i en kostnads- och intäktsanalys är det diskonterade nuvärdet rätt beslutsunderlag och varnar för att använda intäktskostnadskvoten eller internräntan.

Kostnadseffektkvot

Den inkrementella kostnadseffektkvoten (eng. *incremental cost-effectiveness ratio*, *ICER*) är kvoten mellan differensen mellan kostnaderna för de båda alternativen och differensen mellan samma alternativs effektivitet (kostnadseffektkvoten), enligt nedan:

$$ICER = \frac{\text{Cost}_a - \text{Cost}_b}{\text{Effect}_a - \text{Effect}_b}$$

Av denna kvot framgår att flera möjliga resultat är tänkbara. Inom sjukvården kallas en intervention som både är bättre för patienterna och har lägre kostnad för *dominant*. Många interventioner resulterar dock i att effekten ökar med en högre kostnad. Den inkrementella kostnaden per enhet inkrementell effekt är ett mått på den relativa ekonomiska attraktionskraften för en intervention. En intervention med en hög inkrementell kostnadseffektivitet innebär att ett bättre resultat kräver avsevärt högre kostnader och därmed blir mindre attraktivt än en intervention med lägre inkrementell kostnadseffektivitet.

Fördelningseffekter

De ekonomiska effekterna av en viss åtgärd varierar ofta för olika grupper. När man gör en ekonomisk utvärdering bör man därför alltid försöka presentera resultatet ur olika perspektiv. En intervention kan vara mer ekonomiskt fördelaktig än alternativet ur ett samhällsperspektiv men motsatsen för en enskild verksamhet som

tillhandahåller tjänsten. Det är viktig information, särskilt i de situationer där interventioner utformas för populationer som är ekonomiskt sårbara. På motsvarande sätt kan program tyckas vara ekonomiskt fördelaktiga enbart för att de överför resurser från en aktör till en annan. Det är därför viktigt att försöka redovisa fördelningseffekterna för de interventioner som utvärderas. Som exempel kan nämnas att resultatet av en ekonomisk utvärdering av multisystemisk terapi (MST) i Sverige (Olsson, 2009; 2010), redovisades ur såväl MST-deltagarens perspektiv som ur de deltagares som inte fick MST, socialkontorets och samhällets perspektiv. Skillnader i kostnadsresultatet beaktades dessutom både för de individer som erhöll enbart tjänsten och för hela behandlingsgruppen. Slutligen gjordes en sensitivitetsanalys i syfte att bedöma den potentiella effekten på resultatet från förändringar i implementeringen och socialtjänstens remisshantering.

Sensitivitetsanalys

Med sensitivitetsanalysen identifieras och testas viktiga variabler och antaganden i en ekonomisk utvärdering. Valet av känslighetsanalys beror på vilken utvärdering som görs. Det kan råda osäkerhet om hur effektiv en intervention är när det gäller att uppnå en angiven psykosocial effekt eller värdena på parametrarna i modellen. Därför bör man undersöka variablerna och hur de samverkar så att modellen kan hantera alla rimliga tester. Det kan låta motsägelsefullt för många som har lärt sig att planera jämförelser a priori. På grund av de ekonomiska uppgifternas natur och egenskaper bör dock en omfattande sensitivitetsanalys göras för att testa hur robust det ekonomiska resultatet är.

Metoder för hantering av olika typer av osäkerhet (jfr Drummond m.fl., 2005)

Osäkerhet i ekonomiska utvärderingar kan uppstå på grund av:

- oenighet i val av analysmetod
- kraven på data för undersökningen
- behovet av att extrapolera resultatet över tiden, eller från preliminärt till slutligt resultat
- önskan att generalisera resultatet av undersökningen till en annan situation.

Vilken metod som är att föredra för att hantera osäkerhet beror på källan. Metodosäkerhet kan bara hanteras genom känslighetsanalys eller genom att ta fram metodstandarder eller referensfall.

Metoder för hantering av osäkerhet i analyser baserad på individdata:

Typ av osäkerhet	Hantering av osäkerhet
Metod	Referensfall/känslighetsanalys
Urvalsvariation	Statistisk analys
Extrapolering	Modellmetoder
Möjlighet att generalisera/överföra	Känslighetsanalys

Sammanfattning

Det är viktigt att förstå både de ekonomiska effekterna och behandlingseffekterna av interventioner när man arbetar med planering, utformning och beslutsfattande inom socialtjänsten. Som framgår av kapitlet utgör en ekonomisk utvärdering inte bara ett tillägg till effektutvärderingen. En ekonomisk utvärdering förutsätter särskild planering, datainsamling och analys utöver dem som ingår i effektutvärderingen. Några aspekter av den ekonomiska utvärderingen skiljer sig från den metod som används för effektutvärderingen. Avsikten med detta kapitel är att ge läsaren en introduktion till dessa särskilda hänsynstaganden. Den som är intresserad av ämnet uppmanas att ta del av fördjupningslitteraturen. Checklistan kan också vara ett praktiskt verktyg vid planering och utförande av ekonomiska utvärderingar.

Fördjupningslitteratur

Drummond, M. F. & Davies, L. (1991). Economic analysis alongside clinical trials: revisiting the methodological issues. *International Journal of Technology Assessment in Health Care, 7*, 561–573.

Drummond, M. R. & McGuire, A. (Eds.) (2001). *Economic Evaluation in Health Care: Merging Theroy with Practice.* Oxford: Oxford University Press.

Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. J. & Stoddare, G. L. (2005). *Methods for the Economic Evaluation of Health Care Programmes, 3rd Edition*. Oxford, UK: Oxford University Press.

Olsson, T.M. (2012). Economic evaluation as a component of quality effectiveness research: methodological and practical benefits. *Child and Youth Forum, 41*, 137–148.

Referenser

Litteratur

Achenbach, T.M. (1991). *Manual for the Child Behavior Checklist and 1991 profile*. Burlington: University of Vermont, Department of Psychiatry.

Barber, J.A. & Thompson, S.G. (1998). Analysis and interpretation of cost data in randomized controlled trials: review of published studies. *British Medical Journal*, 317, 1195–1200.

Barnett, P.G. (2003). Determination of VA health care costs. *Care research and review, 60*, 124S–141S.

Boardman, A.E., Greenberg, D.H., Vining, A.R. & Weimer, D.L. (2006). *Cost-benefit analysis: Concepts and practice, 3rd Edition.* Upper Saddle River, N.J.: Pearson Prentice Hall.

Briggs, A. H. & Gray, A. M. (1999). Handling uncertainty when performing economic evaluation of healthcare interventions. *Health Technology Assessment, 3*, 1–134.

Brouwer, W.B., van Exel, N., Baltussen, R. & Rutten, F. (2006). A dollar is a dollar is a dollar – or is it? *Value in Health, 9*, 341–347.

Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. J. & Stoddare, G. L. (2005). *Methods for the economic evaluation of health care programmes, 3rd Edition*. Oxford, UK: Oxford University Press.

Gafni, A., Walter, S.D., Birch, S. & Sendi, P. (2008). An opportunity cost approach to sample size calculation in cost-effectiveness analysis, *Health Economics, 17*, 99–107.

Morris, S., Devlin, N. & Parkin, D. (2007). *Economic analysis in health care.* West Sussex, England: John Wiley & Sons, Ltd.

O'Hagan, A. & Stevens, J.W. (2003). Assessing and comparing costs: how robust are the bootstrap and methods based on asymptotic normality? *Health Economics, 12*, 33–49.
Olsson, T.M. (2009). Intervening in youth problem behavior in Sweden: a pragmatic cost analysis of MST from a randomized trial with conduct disordered youth. *International Journal of Social Welfare, 19*, 194–205.
Olsson, T.M. (2010). MST with conduct disordered youth in Sweden: costs and benefits after 2 years. *Research on Social Work Practice, 20*, 561–571 .
Olsson, T.M. (2011). Comparing top-down and bottom-up costing approaches for economic evaluation within social welfare. *European Journal of Health Economics, 12*, 445–453.
Sefton, T., Byford, S., McDaid, D., Hills, J. & Knapp, M. (2992). *Making the most of it: Economic evaluation in the social welfare field.* Layerthorpe, York:Joseph Rowntree Foundation.
Walter, S.D., Garnmi, A. & Birch, S. (2007). Estimation, power, and sample size calculations for stochastic cost and effectiveness analysis. *Pharmoeconomics, 25*, 455–466.
Wonderling, D., Gruen, R. & Black, N. (2005). *Introduction to health economics*. Berkshire, England: Open University Press.

Checklista för bedömning av metodkvalitet vid ekonomisk analys

Punkt		Ja	Nej	Inte tydligt	Ej tillämpligt
Undersökningens upplägg					
1	Frågeställningen formuleras				
2	Frågeställningens ekonomiska betydelse fastställs				
3	Perspektivet för analysen fastställs och motiveras				
4	Den logiska grunden redovisas för att välja vilka kontrollalternativ som ska jämföras				
5	De alternativ som jämförs beskrivs klart och tydligt				
6	Typen av ekonomisk analys som ska tillämpas fastställs				
7	Val av typ av ekonomisk analys motiveras utifrån frågeställningen				
Datainsamling					
8	Källor till effektivitetsuppskattningar anges				
9	Detaljerad information om upplägget och resultatet av effektutvärderingen anges (om den baseras på en enskild studie)				
10	Detaljerad information om metoderna för sammanställning eller metaanalys av estimaten anges (om det bygger på en sammanställning av flera effektivitetsstudier)				
11	De primära resultatmåtten för den ekonomiska utvärderingen anges				
12	Metoder för att värdera nyttan anges				
13	Detaljerad information om de individer som värderingarna baseras på anges				
14	Produktivitetsförändringar (om sådana ingår) redovisas separat				
15	Relevansen av produktivitetsändringarna för frågeställningen avhandlas				
16	Mängden använda resurser redovisas separat från motsvarande enhetskostnader				

17	Metoder för uppskattning av kvantiteter och enhetskostnader beskrivs				
18	Valutauppgifter och prisdata registreras				
19	Detaljerad information om valutan för prisjusteringar vid inflation eller valutaomvandling anges				
20	Detaljerad information om använda modeller anges				
21	Val av modell och viktiga parametrar motiveras				
Analys och tolkning av resultat					
22	Tidsram för kostnader och nytta anges				
23	Diskonteringsräntan anges				
24	Valet av diskonteringsränta motiveras				
25	En förklaring anges om kostnader och intäkter inte har diskonterats				
26	Detaljerad information om statistiska tester och konfidensintervall anges för stokastiska data				
27	Tillvägagångssättet för sensitivitetsanalysen anges				
28	Valet av variabler för sensitivitetsanalysen motiveras				
29	De intervall som variablerna varierar inom motiveras				
30	Relevanta alternativ jämförs				
31	Den inkrementella analysen redovisas				
32	Viktiga resultat redovisas i aggregerad och icke-aggregerad form				
33	Svaret på frågeställningen anges				
34	Slutsatser följer av de data som redovisas				
35	Slutsatserna åtföljs av lämpliga invändningar				

Källa: Co-convenors of the Campbell & Cochrane Economics Methods Group (CCEMG), 2008. The Campbell Collaboration economics methods policy brief, Version 1.0. Norwich, UK: Author.

Kapitel 7 | Laura Ferrer-Wreder & Knut Sundell

Utvärdering av importerade interventioner[1]

Det ökade intresset för en evidensbaserad praktik innebär ett behov av evidensbaserade interventioner (EBI).[2] Det har medfört att interventioner som kvalitetssäkrats i en kultur nu sprids till andra kulturer. Exempelvis handlade 41 procent av avslutade eller pågående svenska effektutvärderingar av psykosociala interventioner sommaren 2010 om importerade EBI (Socialstyrelsen, 2011). Detta kapitel handlar om kulturell anpassning och utvärdering av importerade EBI.

Blandade utfall av effektutvärderingar i nya kulturer

Resultaten från ett antal effektutvärderingar av importerade EBI har delvis bekräftat positiva effekter, delvis producerat resultat där EBI inte skiljer sig från jämförelsegruppen. I tabell 7:1 ges exempel på in-

1 Kapitlet har översatts från engelska till svenska av Åsa Kling, institutionen för psykologi, Uppsala universitet.

2 Med EBI avser vi i detta kapitel interventioner som i minst en experimentell studie med jämförelsegrupp, pre-postmätning och vetenskapligt säkrade datainsamlingsinstrument visat sig vara mer effektiv än sin jämförelse (jfr Flay, Biglan, Boruch, Castro, Gottfredson, Kellam m.fl., 2005).

terventioner som beskrivs som effektiva i US Blueprints Model and Promising Programs (www.colorado.edu/cspv/blueprints) och som utvärderats i randomiserade kontrollerade utvärderingar (RCT) i nya kulturer. För tre av de sex interventionerna i tabell 7:1 finns resultat som inte bekräftat de ursprungliga effekterna.

Det finns åtminstone fyra förklaringar till att de ursprungligen positiva effekterna inte replikerats. En första är att olika forskningsdesigner använts i den ursprungliga och den nya effektutvärderingen. Effektstorlekar är exempelvis ofta större om kontrollgruppen utgörs av väntelista eller placebo jämfört med om den fått ta del av en annan intervention (t.ex. Magill & Ray, 2009; Shadish, 2011). Effekterna är i allmänhet också större i modellutvärderingar där forskaren säkrar hög metodtrohet jämfört med utvärderingar i ordinarie verksamhet (t.ex. Curtis, Ronan & Borduin, 2004; Petrosino & Soydan, 2005). Om bortfall inte hanteras med hjälp av imputering (eng. *Treatment on the treated*, eller *Treatment of treated, TOT*) blir effekterna ofta också högre än om imputering används (eng. *Intention to treat analysis*, *ITT*), där deltagare som avbryter medverkan ingår i analyserna (Wright & Sim, 2003). En andra förklaring är att interventionen inte använts som avsett, med hög behandlingstrohet. En tredje förklaring handlar om dålig kulturell anpassning. Förklaring två och tre kan vara relaterade till varandra; om en intervention inte känns kulturellt relevant kan behandlingstroheten minska. En fjärde förklaring till variation i resultat är att sociodemografiska eller kulturella skillnader modererar effekterna av en EBI i den nya kulturen (se även Sundell & Ferrer-Wreder, under tryckning). Till exempel är det möjligt att interventioner mot cannabismissbruk bland unga är mindre effektiva i Sverige än i ett land som USA eller England där prevalensen är väsentligt högre än i Sverige (ter Bogt, Schmid, Nic Gabhainn, Fotiou & Vollebergh, 2006). För att kunna avgöra vad bristande överensstämmelse i resultat beror på behövs en strategi för hur importerade interventioner ska utvärderas. Annars finns risken att en effektiv EBI förkastas.

Även om kapitlet behandlar kulturell anpassning av importerade

EBI, finns det mycket att lära av forskning om interventioner som anpassats för nya grupper i samma land. Ett stort antal amerikanska studier utvärderar exempelvis dels interventioner som är kulturellt anpassade för speciella etniska minoritetsgrupper i USA, dels effektivitet när interventioner används med andra grupper än de ursprungliga (t.ex. Elliott & Mihalic, 2004).

Tabell 7:1. Exempel på evidensbaserade interventioner som till och med 2011 utvärderats i andra kulturer med replikerade resultat (ja) eller inte replikerade resultat (nej)

Functional Family Therapy

Ja Sverige (Hansson, Cederblad & Höök, 2000)

Incredible Years

Ja England (Gardner, Burton & Klimes, 2006)
Ja Kanada (Taylor, Schmidt, Pepler & Hodgins, 1998)
Ja Norge (Fossum, Mørch, Handegård, Drugli & Larsson, 2009)
Ja Sverige (Axberg & Broberg, under tryckning)

Multidimensional Treatment Foster Care

Ja Sverige (Kyhle Westermark, Hansson & Olsson, 2011)

Multisystemic Therapy

Nej Kanada (Leschied & Cunningham, 2002)
Ja Norge (Ogden & Hagen, 2006)
Nej Sverige (Andrée Löfholm, Olsson, Sundell & Hansson, 2009)

Strengthening Families Program

Nej Sverige (Skärstrand, Sundell & Andreason, under prep.)

Triple-P

Ja Hong Kong (Leung, Sanders, Leung, Mak & Lau, 2003)
Ja Japan (Matsumoto, Sofronoff & Sanders, 2010)
Ja Schweiz (Bodenmann, Cina, Ledermann & Sanders, 2008)
Nej Schweiz (Malti, Ribeaud & Eisner, 2011; Eisner, Nagin, Ribeaud & Malti, 2012)
Ja Tyskland (Hahlweg, Heinrichs, Kuschel & Feldmann, 2007)
Ja USA (Prinz, Sanders, Shapiro, Whitaker & Lutzke, 2009)

Interventionsvetenskap och kultur

Weisz, Sandler, Durlak och Anton (2005) menar att interventioner inriktade på hälsopromotion, positiv utveckling hos barn och unga, prevention och behandling visserligen skiljer sig på en mängd olika sätt men att de också kompletterar varandra. Dessa interventioner har målet att åstadkomma en positiv förändring i människors liv genom planerade interventioner (Weisz m.fl., 2005). De möter också samma problem med att mobilisera resurser i närsamhället, implementera interventioner som avsett, i val av utvärderingsdesign och analys, i vidmakthållande av resultat över tid samt i att sprida effektiva interventioner i stor skala. Kunskap från dessa områden konstituerar det som i dag kallas interventionsvetenskap (Weisz m.fl., 2005).

Inom interventionsvetenskap finns förutsättningar för korsbefruktning av idéer och metoder och många forskare har efterfrågat tvärvetenskaplig forskning som kan minska dagens fragmentiserade kunskapsbas om interventioner (Kurtines, Ferrer-Wreder, Berman, Lorente, Silverman & Montgomery, 2008; Masten, Faden, Zucker & Spear, 2008). Det är nödvändigt om samhället ska lyckas främja positiv utveckling och hälsa, förebygga ohälsa och utanförskap samt behandla ohälsa och sjukdom (Weisz m.fl., 2005).

Enligt Super och Harkness (1999) är kultur ett mångtydigt begrepp som kan mätas på olika sätt betroende på syftet. De definierar kultur som ”en omedelbar, ständigt närvarande verklighet som delas av människor i ett samhälle och som formar och färgar all erfarenhet och allt beteende” (s. 281). Kultur har också en framträdande roll i Weisz och medarbetares modell (2005) för interventionsvetenskap. Enligt dem blir den praktiska nyttan av en intervention större om den utformats för att passa in i kulturen för interventionens målgrupp. Exempel på kulturella skillnader är i vilken utsträckning personer förväntas ta hand om sig själva och den närmaste familjen (individualistiska samhällen) eller om personer integreras i starkt sammanhållna grupper, ofta släkter som fortsätter att skydda dem i utbyte mot okritisk lojalitet (kollektivistiska samhällen).

Kulturell anpassning

Tack vare ett ökat antal artiklar om kulturell anpassning har framsteg gjorts för att förstå kulturens och kontextens roll för behandling av psykisk ohälsa och prevention av problem hos barn och ungdomar (t.ex. Bernal, Jiménez-Chafey & Domenech Rodríguez, 2009; Castro, Barrera & Holleran Steiker, 2010; Huey & Polo, 2008; Poulsen, Vandenhoudt, Wyckoff, Obong'o, Ochura, Njika m.fl., 2010). Fler framsteg är att vänta, bland annat tack vare ett ökat fokus på analyser av mediatorer och moderatorer (t.ex. Bloom & Michalopoulos, 2011; Fairchild & MacKinnon, 2009). Tidigare var det vanligt att forskare genomförde en effektutvärdering med det enda syftet att undersöka interventionens inverkan på ett utfall. Det är fortfarande viktigt, men i dag förväntas också analyser av relationen mellan mediatorer och utfall. Det ger bättre kunskap om validiteten och generaliserbarheten i EBI:s teoretiska förändringsmodeller. Att systematiskt testa en förändringsteori kan vara av stor betydelse för en optimal anpassning av en importerad EBI.

Termen förändringsteori syftar på en interventions tänkta förändringsmekanism. Andra termer är djupstruktur, kärnkomponenter och programteori (Donaldson, 2001; Resnicow, Soler, Braithwaite, Ahluwalia & Butler, 2000). Mer specifikt beskriver förändringsteorin hur en intervention är tänkt att åstadkomma ett önskat utfall. En förändringsteori kan beskrivas i form av en logisk modell eller manual. Den specificerar empiriskt eller teoretiskt härledda relationer mellan interventionens komponenter, processer eller mediatorer samt förändring och utfall (Castro m.fl., 2010).

> ***Interventionens förändringsteori***
> Beskriver hur interventionen är relaterad till förändringsmålet och hur interventionens förändringsmål är relaterade till varandra.

I väl beforskade EBI finns i allmänhet en tydlig förändringsteori. Ofta har förändringsteorierna dessutom åtminstone delvis testats. Exempelvis finns fem sammanlänkade förändringsteorier i Promot-

ing Alternative Thinking Strategies (PATHS; Greenberg, Kusché & Mihalic, 1998). Studier av PATHS har både undersökt interventionens effekter och testat programmets förändringsteorier, exempelvis PATHS beteendeekologiska systemorientering (t.ex. Greenberg, Kusché & Pentz, 2006).

Genom mediator- och moderatoranalyser är det möjligt att testa generaliserbarheten i en förändringsteori för en importerad EBI. Om ungefär samma relation återfinns mellan variabler som medierar förändring och utfall i både den ursprungliga studien och i en replikeringsstudie i en ny kultur ökar sannolikheten att interventionen ger önskad effekt i den nya kulturen.

Hypotetiska relationer mellan teoribaserade mediatorer och utfall (d.v.s. test av interventionens förändringsteori) kan prövas i formativa studier (t.ex. tvärsnittsstudier eller longitudinella studier med kort tidsintervall). En sådan empirisk granskning kan ligga till grund för valet av vilken EBI som ska importeras (Kumpfer, Pinyuchon, Teixeira de Melo & Whiteside, 2008). Om interventionens förändringsteori förefaller relevant för målgruppen i den nya kontexten talar det till interventionens fördel relativt ett alternativ som inte har samma empiriska relevans (Kumpfer m.fl., 2008). Ett sådant

Djup- och ytstruktur
(Resnicow m.fl., 2000)

Djupstruktur, eller en interventions förändringsteori, beskriver hur en intervention är tänkt att fungera för att åstadkomma önskat utfall; vilka de kausala relationerna antas vara. En förändringsteori kan formuleras i form av en logisk modell eller beskrivas i en implementeringsmanual. Den specificerar empiriskt eller teoretiskt härledda relationer mellan interventionens komponenter (mediatorer) och effekter.

Ytstrukturen beskriver hur interventionens budskap och material presenteras. Den bör vara anpassad till det kontext där interventionen ska användas och till livserfarenheter hos programdeltagare. Ytstrukturen är det som ökar acceptans och förståelse för en intervention.

test är en önskvärd men inte tillräcklig förutsättning för en lyckad överföring till en ny kultur.

Utöver formativt test före den egentliga effektutvärderingen kan relationer mellan variabler som förklarar förändring och utfall även undersökas i replikeringsstudier genom analyser av mediator-utfall och moderator-utfall. Dessa analyser tillför fler metodologiska verktyg för att pröva en importerad EBI:s generaliserbarhet.

Behandlingstrohet och anpassning

Trots en ökad kunskap om kulturell anpassning av interventioner återstår flera olösta frågor. En handlar om hur mycket en EBI kan anpassas till en ny kultur utan att det påverkar dess effektivitet. Det finns de som betonar vikten av hög behandlingstrohet och de som framhåller nödvändigheten av att EBI måste anpassas till den nya kulturen (Castro, Barrera & Martinez, 2004; Castro m.fl., 2010; Elliott & Mihalic, 2004). Behandlingstrohet handlar om likhet mellan hur en intervention ska tillämpas enligt ett behandlingsprotokoll eller en manual och hur interventionen faktiskt används i verkligheten (Durlak & DuPre, 2008; Greenberg, Domitrovich, Graczyk & Zins, 2004) – genomfördes interventionen fullständigt och korrekt? Anpassning å andra sidan representerar avvikelse från de föreskrivna riktlinjerna. Sådana ändringar kan vara planerade eller spontana och har ibland syftet att göra interventionen bättre anpassad till en viss implementeringsmiljö (Durlak & DuPre, 2008; Ringwalt, Vincus, Ennett, Johnson & Rohrbach, 2004). Det är mer troligt att en intervention anpassas när den exporteras från ett land till ett annat än när en intervention replikeras i sitt eget hemland (Skärstrand, Larsson & Andréasson, 2008). Enbart att översätta en manual innebär oundvikligen att visst arbete läggs ner på att anpassa interventionen och med det följer vissa kulturella förändringar.

Ett exempel på hur en EBI kan anpassas till en ny kultur är det amerikanska drogpreventionsprogrammet *Strengthening Families Program* (SFP) som har importerats till Sverige under namnet *Steg-för-Steg*. SFP är ett drogpreventionsprogram som riktar sig till elever

och deras föräldrar. Interventionens förändringsteori är att stärka föräldrars kompetens genom färdighetsträning i positiv emotionell kommunikation och adekvat gränssättning utifrån barnets utvecklingsmässiga mognad. Föräldrar får därmed hjälp i att hitta en balans mellan värme och gränssättning, liksom i att upprätthålla och kommunicera förväntningar om att barnet ska avhålla sig från alkohol och droger. Mediatorer för förändring hos de unga inkluderar att utveckla färdigheter, till exempel att stå emot grupptryck och copingstrategier (Kumpfer, Alvarado & Whiteside, 2003; Kumpfer & Turner, 1990; Kumpfer m.fl., 2008).

Kompetens- och färdighetsträning genomförs i sessioner med enbart elever, enbart föräldrar och med elever och föräldrar tillsammans (Kumpfer m.fl., 2008). Rekommenderat antal elev-föräldrapar är åtta till 13 (Skärstrand m.fl., 2008). Interaktiv pedagogik, modellinlärning, beteendeträning och hemuppgifter används i förändringsarbetet. Ett viktigt mål i SFP:s förändringsteori är fördröjd alkoholdebut (Spoth, Trudeau, Guyll & Shin, 2012). SFP finns för universella, selektiva och indikerade målgrupper (Kumpfer m.fl., 2008). Den version som implementerades och utvärderades i Sverige riktade sig till en universell målgrupp och kallas SFP 10–14. SFP 10–14 består av sju sessioner och en ”boostersession” efter sex och tolv månader (Kumpfer, Molgaard & Spoth, 1996).

Den amerikanska evidensen för SFP 10–14 omfattar bland annat en randomiserad kontrollerad studie som testade effekterna av SFP 10–14 med företrädesvis vita tonåringar och deras familjer i mindre orter i Iowa (Spoth, Redmond & Lepper, 1999). Ungdomarna följdes fram till att de var i ung vuxenålder. Kortsiktiga resultat visar på en fördröjd alkoholdebut hos ungdomarna som deltog i SFP. Tio år efter interventionen fanns positiva indirekta, medierade effekter. Spoth och medarbetare (2012) drog slutsatsen att fördröjd alkoholdebut i tonåren hos unga som deltagit i SFP 10–14 medierade SFP:s positiva effekter på problematiskt drog- och alkoholbruk i ung vuxenålder. De positiva effekterna kunde med andra ord hänföras till att SFP 10–14 fördröjt alkoholdebuten under tonåren.

Under 2008 publicerade SFP:s programutvecklare rekommendationer för internationell spridning av SFP (Kumpfer m.fl., 2008). Dessa rekommendationer var inte tillgängliga vid tiden för den svenska SFP-studien, men deras modell för kulturell anpassning av SFP överensstämmer i allt väsentligt med anpassningsprocessen i den svenska SFP 10–14-studien. Ett krav från SFP:s programkonstruktörer är att förändringsteori (d.v.s. interventionens djupstruktur) och behandlingsdos inte får ändras när interventionen importeras till andra kulturer. Däremot är ytstrukturella förändringar rekommenderade, till exempel att material och budskap ska omarbetas så att de är kulturellt acceptabla och relevanta för deltagarna (Kumpfer m.fl., 2008).

Den svenska SFP-studien startade 2001 med en förstudie för att anpassa interventionen till svenska förhållanden (Skärstrand, Bränström, Sundell, Källmén & Andréasson, 2009). Därefter genomförde de svenska forskarna den kulturella anpassningsprocessen i samarbete med programutvecklarna (Skärstrand m.fl., 2008). Först gjordes ett antal ytstrukturella förändringar i programmet. Allt material översattes från engelska till svenska och nya videovinjetter spelades in med samma innehåll och syfte som de amerikanska, men med en svensk publik i åtanke (Skärstrand m.fl., 2008). Ändringarna baserades på återkoppling från referensgrupper, en pilotstudie i mindre skala av SFP samt intervjuer med tänkbara programdeltagare. Då arbetsschemat för svenska lärare hindrade att de deltog i sessioner efter skoldagens slut, medverkade de endast i sessioner som hölls under skoltid. SFP implementeras vanligtvis med separata men samtidiga sessioner som följs av en session där både de unga och föräldrar deltar. Eftersom föräldrasessioner under dagtid skulle krocka med de flesta föräldrars arbeten, var det nödvändigt att dela upp tiden mellan de ungas och föräldrarnas sessioner så att elevsessionerna gavs under skoltid och föräldrasessionerna gavs på kvällen samma dag. De gemensamma sessionerna komprimerades till två 2-timmarssessioner. För de senare sessionerna användes andra ledare.

Andra ändringar var att boostersessionerna integrerades i pro-

grammet i stället för att läggas efter den egentliga interventionen. Det var också omkring dubbelt så många deltagare i de svenska SFP-sessionerna utan att antalet gruppledare ökades. Programinnehållet i den svenska versionen följde manualen, med undantag för en extra session som specifikt ägnades åt att förebygga användning av droger och alkohol. Dessutom utelämnades vissa aktiviteter som bedömdes som ogörliga i Sverige. Dit hörde att föräldrarna skulle avsluta varje session med att ställa sig i cirkel och hålla varandra i händerna, något som betraktades som alltför religiöst eller ”amerikanskt”.

Den svenska effektutvärderingen genomfördes som en klusterrandomiserad verksamhetsbaserad utvärdering av 19 grundskolor och närmare 700 elever. Skolorna som lottades till kontrollgruppen fick inget extra stöd. SFP implementerades med hög behandlingstrohet och utvärderingen pågick mellan 2003 och 2006. Resultaten visar inte på några positiva effekter för SFP 10–14 relativt kontrollgruppen på vårterminen i årskurs 9 (Skärstrand, Caria, Sundell & Andréasson, 2012).

Även om anpassningsprocessen genomfördes omsorgsfullt är det inte möjligt att avgöra om avsaknaden av effekter för SFP 10–14 beror på bristande anpassning, att djupstrukturen förvanskats (t.ex. att de gemensamma sessionerna komprimerades till två) eller på att interventionen är ineffektiv i en svensk kontext. Svenska föräldrar är kanske generellt sett mer restriktiva till tidig alkoholdebut än i USA? Kan de blygsamma ändringarna i ytstrukturen eller programformatet ha gjort att interventionen inte förmådde generera de förväntade positiva effekterna? Mot bakgrund av den här typen av frågor har Castro och medarbetare (2010, s. 213) uppmärksammat behovet av en ny generation forskning om kulturell anpassning. En konsekvens av det perspektivet är att kultur, och därmed behandlingstrohet och anpassning, bör beaktas under hela processen av programutveckling och utvärdering. Kultur kan inte längre bara vara en fråga under urvalsprocessen av en importerad EBI eller för en förstudie. En sådan ny forskningsstrategi skulle möjliggöra större kunskap om optimal kulturell anpassning liksom fördjupning av begrepp och generaliserbarheten i EBI:s förändringsteorier.

Protokollet för Planerad Interventions-Anpassning (PIA)

Protokoll för Planerad Interventions-Anpassning (PIA) är ett av flera teoretiska verktyg (Ferrer-Wreder, Sundell & Mansoory, 2012) för att anpassa och utvärdera importerade EBI. PIA baseras på flera befintliga teorier och modeller och är utformat för att med vetenskaplig metodik avgöra behov och omfattning av anpassningar av en importerad EBI till en ny kultur.

Skillnaden mellan interventioners djupstruktur och ytstruktur är en grundläggande utgångspunkt (Resnicow m.fl., 2000)[3] och PIA-protokollet består av två faser. Den första fasen består av ett antal förberedande moment för att skapa underlag för kulturell anpassning, exempelvis fokusgrupper eller intervjuer för att göra ytstrukturella ändringar, översättning och tillbakaöversättning (tabell 7:2). Andra moment i denna första fas är mindre vanliga i forskning om kulturell anpassning. Dit hör att testa interventionens djupstruktur genom att undersöka sambandet mellan de tänkta mediatorerna och utfallsvariablerna i en tvärsnittstudie. Med hjälp av informationen i fas 1 skapas en minimalt anpassad EBI (enbart ytstrukturella förändringar) och en anpassad version som kan innehålla djupstrukturella förändringar.

Den andra fasen i PIA-protokollet handlar om att med hjälp av experimentell forskningsdesign testa den minimalt anpassade EBI och den anpassade versionen med en kontrollgrupp (t.ex. TAU eller väntelista) genom för-, efter- och eventuellt uppföljningsmätning. Resultaten bör kompletteras med mediator- och moderator-analyser. Vid behov kan därefter ytterligare revideringar och anpassningar göras (tabell 7:3).

Fördelen med PIA-protokollet är att det ger en möjlighet att särskilja effekter från ytstrukturella och djupstrukturella förändringar och därmed också skapa bättre förståelse för kulturens betydelse.

3 Se även Sundell & Ferrer-Wreder, 2012.

Den empiriska evidensen för PIA-protokollet återstår att etableras, men liknar i allt väsentligt andra modeller för området kulturell anpassning.

Sammanfattning

Allt fler effektutvärderingar misslyckas med att replikera ursprungsresultaten när de utvärderas i en ny kontext. Det kan bero på skillnader i forskningsdesign, låg behandlingstrohet, dålig kulturell anpassning eller att interventionen verkligen inte är effektiv i den nya kontexten. Därför finns behov av nya strategier och teorier för hur importerade EBI ska utvärderas för att öka sannolikheten att hitta optimala nivåer av anpassning för interventioner. Lovande strategier för framtida forskning inom detta område inkluderar deskriptiva formativa studier till stöd för anpassningar liksom användandet av inslag från experimentell design i effektivitetsstudier för att kunna jämföra graden av anpassning hos interventionen med nyttan av interventionen (Castro m.fl., 2010).

Tabell 7:2. Protokoll för Planerad Interventions-Anpassning (PIA) – Fas I

A. Ett forum skapas för samarbete mellan programutvecklare och användare i den nya kulturella kontexten. En idealisk intressentgrupp består av forskare, professionella, beslutsfattare och representanter från den tänkta målgruppen för interventionen.
B. Mellan 200 och 400 deltagare rekryteras från interventionens målgrupp. De får delta i en serie formativa studier som ger underlag för anpassningen av den importerade EBI. Både kvantitativa och kvalitativa analyser kan vara aktuella.
C. I samråd med programutvecklarna översätts (vid behov även "tillbakaöversättning") det nya programmets material och effektmått (frågeformulär, observationsprotokoll och annat som mäter interventionens effekter). Dessa instrument ska sedan användas som utfallsmått i effektutvärderingen.
D. De rekryterade deltagarna delas sedan i två ungefär lika stora grupper. Den första gruppen får testa de översatta effektmåtten i en tvärsnitts-

studie. Resultatet analyseras psykometriskt (inklusive reliabilitets- och validitetsanalyser) och revideras vid behov.

E. Den andra hälften av de rekryterade deltagarna testas med de reviderade utfallsmåtten i punkt D i en tvärsnittsdesign eller i en longitudinell studie med kort tidsintervall. Därefter analyseras resultatet statistiskt (lämpligen med mediatoranalys) för att undersöka djupstrukturen i den importerade EBI. Från dessa deltagare väljs 15–20 ut för att medverka i en fokusgrupp om interventionens ytstruktur (material, aktiviteter) för att bedöma om den är godtagbar.
F. Användare och programutvecklare använder resultaten för att utveckla en anpassad version av den importerade EBI. Den anpassade versionen testas sedan i en mindre pilotstudie som ger möjlighet till ytterligare förändringar i utformningen.

Tabell 7:3. Protokoll för Planerad Interventions-Anpassning (PIA) – Fas II

Fas I av PIA-protokollet bör ha resulterat i en minimalt anpassad version av den importerade interventionen med endast språkliga och ytstrukturella förändringar och en fullt anpassad version. Båda versionerna av interventionen utformas i samarbete mellan EBI:s programutvecklare och användare.

G. En ny grupp deltagare rekryteras till en effektutvärdering, baserat på powerberäkning. Den minimalt anpassade versionen jämförs med den fullt anpassade versionen och en jämförelsegrupp (t.ex. standardbehandling eller väntelista), det vill säga sammanlagt tre grupper. Det här steget har rekommenderats i forskningslitteraturen om kulturell anpassning, men få studier har tillämpat det (Castro m.fl., 2010).
H. Därefter genomförs mediator-moderator-analyser för att undersöka vilka effekter som de två anpassade interventionerna ger jämfört med jämförelsegruppen. Dessa analyser utgör samtidigt ett andra test av den importerade EBI:s förändringsteori i den nya kontexten.
I. Vid behov genomförs ytterligare revideringar och anpassningar av den importerade EBI i samarbete mellan programutvecklare och användare. Ytterligare effektutvärderingar kan bli aktuella, framför allt om resultaten är positiva (t.ex. vid bred implementering av interventionen).

Fördjupningslitteratur

Castro, F. G., Barrera, M. & Holleran Steiker, L. K. (2010). Issues and challenges in the design of culturally adapted evidence-based interventions. *Annual Review of Clinical Psychology, 6*, 213-239.

Durlak, J. & DuPre, E. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41*, 327-350.

Kumpfer, K. L., Pinyuchon, M., Teixeira de Melo, A. & Whiteside, H. O. (2008). Cultural adaptation process for international dissemination of the Strengthening Families Program. *Evaluation & the Health Professions, 31*, 226–239.

Resnicow, K., Soler, R., Braithwaite, R. L., Ahluwalia, J. S. & Butler, J. (2000). Cultural sensitivity in substance use prevention. *Journal of Community Psychology, 28*(3), 271-290.

Sundell, K. & Ferrer-Wreder, L. (under tryckning). The Transportability of Empirically-Supported Interventions. Aron Shlonsky och Rami Benbenishty (Red.), *Using evidence in child welfare*. Oxford University Press.

Referenser

Andrée Löfholm, C., Olsson, T., Sundell, K. & Hansson, K. (2009). Multisystemic Therapy with conduct disordered youth: Stability of treatment outcomes two years after intake. *Evidence & Policy, 5,* 373–397.

Axberg, U. & Broberg, A. G. (under tryckning). Evaluation of 'The Incredible Years' in Sweden: The transferability of an American parent-training programme to Sweden. *Scandinavian Journal of Psychology.*

Bernal, G., Jiménez-Chafey, M. I. & Domenech Rodríguez, M. M. (2009). Cultural adaptation of treatments: A resource for considering culture in evidence-based practice. *Professional Psychology: Research and Practice, 40*, 361–368.

Bloom, H. S. & Michalopoulos, C. (2011). When is the story in the subgroups? Strategies for interpreting and reporting intervention effects for subgroups. *Prevention Science* (först publicerat on-line 29 januari 2011: DOI 10.1007/s11121-010-0198-x:

Bodenmann, G., Cina, A., Ledermann, T. & Sanders, M. R. (2008). The efficacy of the Triple P-Positive Parenting Program in improving parenting and child behavior: A comparison with two other treatment conditions. *Behaviour Research and Therapy, 46,* 411–427.

ter Bogt, T., Schmid, H., Nic Gabhainn, S., Fotiou, A. & Vollebergh, W. (2006). Economic and cultural correlates of cannabis use among mid-adolescents in

31 countries. *Addiction, 101,* 241–251.
Castro, F. G., Barrera, M. & Martinez, C. R. (2004). The cultural adaptation of prevention interventions: Resolving tensions between fidelity and fit. *Prevention Science, 5*, 41–45.
Castro, F. G., Barrera, M. & Holleran Steiker, L. K. (2010). Issues and challenges in the design of culturally adapted evidence-based interventions. *Annual Review of Clinical Psychology, 6*, 213–239.
Curtis, N. M., Ronan, K. R. & Borduin, C. M. (2004). Multisystemic treatment: a meta-analysis of outcome studies. *Journal of Family Psychology, 18*, 411–419.
Donaldson, S. I. (2001). Overcoming our negative reputation: Evaluation becomes known as a helping profession. *American Journal of Evaluation, 22*, 355–361.
Durlak, J. & DuPre, E. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41*, 327–350.
Eisner, M., Nagin, D., Ribeaud, D. & Malti, T. (2012). Effects of a universal parenting program for highly adherent parents: A propensity score matching approach. Prevention science. (Först publicerat 9 januari 2012 on-line: DOI 10.1007/s11121-011-0266-x).
Elliott, D. S. & Mihalic, S. (2004). Issues in Disseminating and Replicating Effective Prevention Programs. *Prevention Science, 5*, 47–53.
Fairchild, A. J. & MacKinnon, D. P. (2009). A general model for testing mediation and moderation effects. *Prevention Science, 10*, 87–99.
Ferrer-Wreder, L., Sundell, K. & Mansoory, S. (2012). Tinkering with Perfection: Theory Development in the Intervention Cultural Adaptation Field, *Child & Youth Care Forum*, *41*, 149–171.
Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., m.fl. (2005). Standards of evidence. Criteria for efficacy, effectiveness and dissemination. *Preventions science, 6,* 151–175.
Fossum, S., Mørch, W-T., Handegård, B. H., Drugli, M. B. & Larsson, B. (2009). Parent training for young Norwegian children with ODD and CD problems: Predictors and mediators of treatment outcome. *Scandinavian Journal of Psychology, 50* 173–181.
Gardner, F., Burton, J. & Klimes, I. (2006). Randomised controlled trial of a parenting intervention in the voluntary sector for reducing child conduct problems: Outcomes and mechanisms of change. *Journal of Child Psychology and Psychiatry, 47,* 1123–1132.
Greenberg, M. T., Domitrovich, C. E., Graczyk, P. A. & Zins, J. E. (2004). *The study of implementation in school-based preventive interventions: Theory, research, and practice*. Washington, DC: U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Center for Mental Health Services. Final project report.
Greenberg, M.T., Kusché, C. & Mihalic, S.F. (1998). Blueprints for Violence

Prevention, Book Ten: Promoting Alternative Thinking Strategies (PATHS). Boulder, CO: Center for the Study and Prevention of Violence.
Greenberg, M. T. & Walls, C. T. (2003). Examining the role of implementation quality in school-based prevention using the PATHS curriculum. *Prevention Science, 4*, 55–63.
Greenberg, M. T., Kusché, C. A. & Pentz, M. A. (2006). The mediational role of neurocognition in the behavioral outcomes of a social-emotional prevention program in elementary school students: Effects of the PATHS curriculum.
Hahlweg, K., Heinrichs, N., Kuschel, A. & Feldmann, M. (2007). Therapist-assisted, self-administered bibliotherapy to enhance parental competence short- and long-term effects. *Behavior Modification 32*, 659–681.
Hansson, K., Cederblad, M. & Höök, B. (2000) Funktionell familjeterapi. *Socialvetenskaplig tidskrift, 3*, 231–243.
Huey, S. J. & Polo, A. J. (2008). Evidence-based psychosocial treatments for ethnic minority youth. *Journal of Clinical Child and Adolescent Psychology, 37*, 262–301. doi:10.1080/15374410701820174 Psychology, City University of New York, The City College of New York
Kumpfer, K. L., Alvarado, R. & Whiteside, H. O. (2003). Family-based interventions for substance use and misuse prevention. *Substance use & Misuse, 38*, 1759–1787.
Kumpfer, K. L., Molgaard, V. & Spoth, R. (1996). The Strengthening Families Program for the prevention of delinquency and drug use. In R. D. Peters & R. J. McMahon (Eds.), *Preventing childhood disorders, substance abuse, and delinquency* (pp. 241-267). Thousand Oaks, CA: Sage.
Kumpfer, K. L., Pinyuchon, M., Teixeira de Melo, A. & Whiteside, H. O. (2008). Cultural adaptation process for international dissemination of the Strengthening Families Program. *Evaluation & the Health Professions, 31*, 226–239.
Kumpfer, K. L. & Turner, C. W. (1990). The social ecology model of adolescent substance abuse: Implications for prevention. *International Journal of the Addictions, 25*, 435–463.
Kurtines, W. M., Ferrer-Wreder, L., Berman, S. L., Lorente, C. C., Silverman, W. K. & Montgomery, M. J. (2008). Promoting positive youth development – New directions in developmental theory, methods, and research. *Journal of Adolescent Research, 23*, 233–244.
Kyhle Westermark, P., Hansson, K. & Olsson, M. (2011). Multidimensional treatment foster care (MTFC): Results from an independent replication. *Journal of Family Therapy, 33,* 20–41.
Leschied, A.W. & Cunningham, A. (2002). *Seeking effective interventions for serious young offenders – Interim results of a fouryear randomized study of Multisystemic Therapy in Ontario,Canada. Centre for Children & Families in the Justice System*. Nedladdad 2010-02-15 från www.lfcc.on.ca/mst_final_results.html
Leung, C., Sanders, M. R., Leung, S., Mak, R. & Lau, J. (2003). An Outcome Evaluation of the Implementation of the Triple P-Positive Parenting Program

in Hong Kong. *Family Process, 42,* 531–544.
Magill, M. & Ray, L. A. (2009). Cognitive-behavioral treatment with adult alcohol and illicit drug users: a meta-analysis of randomized controlled trials. *Journal of Studies on Alcohol and Drugs, 80,* 516–527.
Malti, T., Ribeaud, D. & Eisner, M. (2011). The effectiveness of two universal preventive interventions in reducing children's externalizing behavior: A cluster-randomized controlled trial. *Journal of Clinical Child and Adolescent Psychology, 40,* 677–692.
Matsumoto, Y., Sofronoff, K. & Sanders, M. R. (2010). Investigation of the effectiveness and social validity of the Triple P Positive Parenting Program in Japanese society. *Journal of Family Psychology, 24*, 87–91.
Masten, A. S., Faden, V. B., Zucker, R. A. & Spear, L. P. (2008). Underage drinking: A developmental framework. *Pediatrics, 121, Supplement 4,* S235–S251.
Ogden, T. & Hagen, K. A. (2006). Multisystemic therapy of serious behavior problems in youth: Sustainability of therapy effectiveness two years after intake. *Child and Adolescent Mental Health, 11*, 142–149.
Petrosino, A. & Soydan, H. (2005). The impact of program developers as evaluators on criminal recidivism: Results from a meta-analysis of experimental and quasi-experimental research. *Journal of Experimental Criminology, 1*, 435–450.
Poulsen, M. N., Vandenhoudt, H., Wyckoff, S. C., Obong'o, C. O., Ochura, J., Njika, G., m.fl. (2010). Cultural adaptation of a U.S. evidence-based parenting intervention for rural western Kenya: From parents matter! To families matter! *AIDS Education and Prevention, 22*, 273–285.
Prinz, R. J., Sanders, M. R., Shapiro, C. J., Whitaker, D. J. & Lutzke, J. R. (2009). Population-Based Prevention of Child Maltreatment: The U.S. Triple P System Population Trial. *Prevention Science, 10,* 1–12.
Redmond, C. (2009). Universal intervention effects on substance use among young adults mediated by delayed adolescent substance initiation. *Journal of Consulting and Clinical Psychology 77*, 620–632.
Resnicow, K., Soler, R., Braithwaite, R. L., Ahluwalia, J. S. & Butler, J. (2000). Cultural sensitivity in substance use prevention. *Journal of Community Psychology, 28*, 271–290.
Ringwalt, C. L., Vincus, A., Ennett, S., Johnson, R. & Rohrbach, L. A. (2004). Reasons for teachers' adaptation of substance use prevention curricula in schools with non-white student populations. *Prevention Science, 5*, 61–67.
Shadish, W. R. (2011). Randomized Controlled Studies and Alternative Designs in Outcome Studies: Challenges and Opportunities. *Research on social work practice, 21,* 636–643.
Skärstrand, E., Bränström, R., Sundell, K., Källmén, H. & Andréasson, S. (2009). Parental participation and retention in an alcohol preventive family-focused programme. *Health Education, 109*, 384–395.
Skärstrand, E., Caria, M. P., Sundell, K. & Andréasson, S. (2012). Evaluation of a Swedish version of the Strengthening Families Programme: Three-year out-

comes of a cluster randomized trial. *Scandinavian Journal of Public Health*.
Skärstrand, E., Larsson, J. & Andréasson, S. (2008). Cultural adaptation of the Strengthening Families Programme to a Swedish setting. *Health Education, 108,* 287–300.
Socialstyrelsen (2011). *Svensk och internationell forskning om sociala interventioners effekter* Stockholm: Socialstyrelsen.
Spoth, R., Redmond, C. & Lepper, H. (1999). Alcohol initiation outcomes of universal family-focused preventive interventions: One-and two-year follow-ups of a controlled study. *Journal of Studies on Alcohol, Suppl. 13*, 103–111.
Spoth, R. L., Trudeau, L. S., Guyll, M. & Shin, C. (2012). Benefits of universal intervention effects on youth protective shield 10 years after baseline. *Journal of Adolescent Health, 50*, 414–417.
Sundell, K. & Ferrer-Wreder, L. (under tryckning). The Transportability of Empirically-Supported Interventions. I Aron Shlonsky och Rami Benbenishty (Red.), *Using evidence in child welfare*. Oxford University Press.
Super, C. M. & Harkness, S. (1999). The environment as culture in developmental research. In S. L. Friedman & T. D. Wachs (Eds.), *Measuring environment across the life span* (pp. 279–323). Washington, DC: American Psychological Association.
Taylor, T. K., Schmidt, F., Pepler, D. & Hodgins, C. (1998). A comparison of eclectic treatment with webster-stratton's parents and children series in a children's mental health center: A randomized controlled trial. *Behavior Therapy, 29,* 221–240.
Weisz, J. R., Sandler, I. N., Durlak, J. A. & Anton, B. S. (2005). Promoting and protecting youth mental health through evidence-based prevention and treatment. *American Psychologist, 60*, 628–648.
Wright C. C. & Sim, J. (2003). Intention to treat approach to data from randomized controlled trials: A sensitivity analysis. *Journal of Clinical Epidemiology, 56,* 533–842.

Kapitel 8 | Maria Bodin

Urval och rekrytering av undersökningsgrupp

Det här kapitlet handlar om urval och rekrytering av undersökningsgrupp till en effektutvärdering. Medan undersökningsgruppens egenskaper bestäms av studiens syfte och mål och därför till stor del är givna på förhand, är själva rekryteringen av studiedeltagare ofta förknippad med problem. Det uppskattas att mindre än hälften av påbörjade effektutvärderingar uppnår sitt rekryteringsmål (Treweek, Pitkethly, Cook, Kjeldstrøm, Taskila, Johansen m.fl., 2010). Ofta behöver rekryteringsfasen till randomiserade studier förlängas, med ökade finansieringskrav som följd. En alltför liten undersökningsgrupp försämrar möjligheten att besvara den primära forskningsfrågan om interventionens effekter, vilket i slutändan får konsekvenser för resultatens publicerbarhet. Forskaren kan vid studiens slut även finna sig i det etiska dilemmat att ha exponerat deltagare för en intervention med okänd effekt, utan att kunna avgöra om den gjort mer nytta än skada. Det är därför viktigt att lägga tid och resurser på att planera rekryteringen av undersökningsgrupp till en effektstudie.

De praktiska momenten i en rekryteringsprocess sammanfattas i Figur 8:1 (jfr Berger, Begun & Otto-Salaj, 2009). Även om de olika momenten tidsmässigt följer efter varandra vid rekryteringen av enskilda individer, behöver samtliga moment beaktas och planeras innan rekryteringen börjar. Planeringen av de initiala aktiviteterna

Figur 8:1. Praktiska moment i rekryteringsprocessen inom effektutvärderingar.

i ruta 1 är exempelvis beroende av urvalskriterierna i steg 3. De olika momenten påverkar och påverkas också av faktorer som har att göra med den studerade interventionen, populationen och det sammanhang där studien äger rum. Kapitlet behandlar vanliga svårigheter i anslutning till urval och rekrytering av undersökningsgrupper för dessa fyra moment. De delar i boken som avser etik (kapitel 3) och forskningsdesigner (kapitel 4) rekommenderas som förberedelse för läsare med mindre erfarenhet av området.

Urval av undersökningsgrupp

De rekryteringsytor och urvalskriterier som används inom en effektutvärdering avgör hur relevanta resultaten blir för andra sammanhang och hur väl man kan besvara forskningsfrågan. För att optimera en effektutvärderings externa och interna validitet skulle forskaren först behöva dra ett slumpmässigt urval av individer eller enheter från den aktuella populationen, och därefter fördela dem slumpmässigt (randomisera) till interventions- och kontrollgrupper. Medan det slumpmässiga urvalet maximerar sannolikheten att undersökningsgruppen är representativ för målpopulationen som helhet och därmed bidrar till studiens externa validitet, maximerar randomiseringen möjligheten till god intern validitet (jfr kapitel 4). En slumpmässigt utvald undersökningsgrupp är emellertid en praktisk omöjlighet i interventionsforskning, eftersom individer själva bestämmer om de vill delta i en studie eller ej. Ett slumpmässigt urval kräver dessutom att hela målpopulationen är sökbar, det vill säga registrerad. När insatsen är en behandling eller riktar sig till riskgrupper, vilket ofta är fallet inom

psykosociala verksamhetsfält, är målpopulationen i sin helhet okänd. Under dessa förutsättningar blir det inte resursmässigt försvarbart att kontakta ett slumpmässigt urval av personer ur befolkningen för förfrågan om deltagande, screening samt eventuell vidare bedömning. I stället behöver forskaren använda sig av kontaktytor där chanserna att nå målgruppen är större, exempelvis via psykiatriska mottagningar, socialkontor eller primärvårdsmottagningar. Detta begränsar undersökningsgruppen till personer som bland annat är mer hjälpsökande än målpopulationen som helhet. Å andra sidan ger rekrytering via vårdinrättningar en mer kliniskt relevant undersökningsgrupp än exempelvis via annonsering i massmedia, där man ofta når personer med lägre problemtyngd. I effektutvärderingar har den interna validiteten vanligtvis företräde framför den externa validiteten (jfr kapitel 4). Det är med andra ord viktigare att man på ett säkert sätt kan studera förändring över tid och hänföra den till insatsen än att studiens resultat går att generalisera till hela målpopulationen. Man bör dock alltid sträva efter att rekrytera en representativ undersökningsgrupp.

> ***Långt utdragen rekryteringsperiod***
> Beardslees familjeintervention är en sekundärpreventiv insats som syftar till att hjälpa deprimerade vuxna att öppna upp en kommunikation kring depressionen med sina barn. På lång sikt vill man på detta vis förebygga att barnen själva insjuknar. Under en finsk utvärdering av insatsen tog det tre år för ett tjugotal behandlingsenheter att rekrytera 119 familjer till studien. Rekryteringen försvårades av att man riktade sig till kliniskt deprimerade människor och att insatsen syftade till att förebygga problem som ännu inte uppstått (Solantaus, Paavonen, Toikka & Punamäki, 2010).

Inklusions- och exklusionskriterier används på individnivå för att ringa in en undersökningsgrupp som utgör målgruppen för den aktuella interventionen (d.v.s. där deltagarna delar det tillstånd eller problem som insatsen avser att förändra) och som är så homogen att den tillåter testning av studiehypotesen. Samtidigt bör gruppen vara så varierad att den är kliniskt relevant och tillåter analyser avseende subgrupper, om sådana planeras. Nedan ses ett exempel på

urvalskriterier som använts i utvärderingar av multisystemisk terapi för ungdomar med allvarliga beteendeproblem i Norge och Sverige (Ogden & Halliday-Boykins, 2004; Sundell, Hansson, Andrée Löfholm, Olsson, Gustle & Kadesjö, 2008):

> The target group was defined as youths of ages 12–17 years who fulfilled the criteria for a clinical diagnosis of conduct disorder according to the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., Text Revision [DSM–IV–TR]; American Psychiatric Association, 2000) and whose parent(s) or parent surrogate(s) were motivated to engage in an intervention. Exclusion criteria were (a) ongoing treatment by another provider; (b) substance abuse without other antisocial behavior; (c) sexual offending; (d) autism, acute psychosis, or imminent risk of suicide; and (e) the presence of the youth in the home constituting a serious risk to the youth or the family.

Vanligtvis eftersträvar man en mer homogen undersökningsgrupp inom en modellutvärdering (eng. *efficacy study*) än inom en utvärdering i ordinarie verksamhet (eng. *effectiveness*). Faktorer som ofta exkluderar medverkan är sådana som försvårar tolkning av resultaten (t.ex. annan pågående behandling), gör insatsen olämplig (t.ex. graviditet vid läkemedelsstudier) eller som avsevärt försvårar deltagande i studien (t.ex. en förestående flytt). När man rekryterar i klinisk verksamhet och träffar klienten eller patienten vid flera tillfällen kan en ny bedömning vid ett senare tillfälle ge en chans att inkludera deltagare som tidigare inte varit möjliga (Berger m.fl., 2009). Personer som tidigare exkluderats på grund av exempelvis en annan pågående behandling kan då inkluderas om denna avslutats.

Det är viktigt att redogöra för hur man bedömt ett urvalskriterium. Används skalor bör psykometriska data för dessa anges. I exemplet ovan skulle forskarna ha kunnat beskriva hur bedömningen av DSM-diagnos gick till. Det är även viktigt att uttömmande beskriva de kontaktytor och urvalskriterier som använts vid rekryteringen så att läsare ges möjlighet att bedöma resultatens generaliserbarhet.

Detta är dessutom ett krav för publicering i de tidskrifter som antagit CONSORT-gruppens riktlinjer för rapportering av randomiserat kontrollerade utvärderingar (RCT) (jfr kapitel 18).

Faktorer som påverkar rekrytering

Studiedesign, interventionen och målgruppen påverkar rekryteringstakten. Forskningsöversikter visar att möjligheterna att uppnå det planerade antalet undersökningspersoner förbättras då forskaren undersöker den förväntade inkluderingstakten (t.ex. tillgång till gruppen, förekomst, det förväntade intresset) genom en pilotrekrytering. Erfarenheter tyder på att professionella som möter målgruppen i sitt arbete tenderar att överskatta tillgången på möjliga deltagare. Många har också en överdriven rädsla att klienterna ska tacka nej till att medverka i randomiserade studier, vilket kan medföra att de undviker att remittera dem till forskaren. Genom en pilotrekrytering ges en mer realistisk bild av vad som kan förväntas. Med dess hjälp kan en uppskattning göras av hur många personer eller enheter som behöver kontaktas, informeras och tillfrågas om samtycke samt screenas för att en person ska kunna inkluderas i studien. Här kan även möjliga fallgropar i processen identifieras och åtgärdas för att det slutliga rekryteringsarbetet ska avlöpa så smidigt som möjligt. Informationen kan också användas som budgetunderlag avseende resursåtgång för rekryteringsfasen. Om interventionen redan är utvärderad i en kontrollerad studie kan den ansvarige forskaren säkerligen bistå med information och erfarenheter.

> ***Rekryteringsproblem ändrade designen***
> Inför en randomiserad studie jämförande två skolbaserade preventionsprogram med en kontrollgrupp beräknades att 60 skolor behövdes för att kunna hitta effekter av den storlek som är vanlig vid liknande insatser. Forskarna lyckades dock endast rekrytera 40 skolor. Man valde då att ta bort en av interventionerna och i stället genomfördes en studie med endast ett preventionsprogram och en kontrollgrupp, för vilken 40 skolor bedömdes ge tillräcklig statistisk power (Bodin & Strandberg, 2011).

Forskaren behöver även skaffa sig en bild av de organisationsfaktorer som påverkar rekryteringsprocessen. Har till exempel en enskild skola mandat att besluta om medverkan i en studie eller tas sådana beslut på en annan nivå? Finns det faktorer som kan hota eller i värsta fall avsluta ett påbörjat rekryteringsarbete? Några exempel är att en verksamhet omorganiseras, går i konkurs eller får en ny ledning som upphäver den tidigare ledningens beslut om att delta. Även om sådana situationer ofta ligger utanför den enskilde forskarens kontroll, kan förebyggande åtgärder vidtas. Forskaren kan till exempel klargöra att utvärderingen endast startar mot löfte om att få fortsätta till dess att rekryteringsmålet uppnåtts. Även andra typer av överenskommelser kan göras, exempelvis att inte exponera kontrollgruppen för den studerade insatsen innan uppföljningsmätningarna genomförts. Även då sådana överenskommelser inte kan vara formellt bindande är de moraliskt förpliktigande. Vid personalbyten behöver forskaren komma ihåg att förnya dessa med den nya verksamhetsansvarige och inte lita på att informationen förs över av den avgående.

Studiedesign

Rekrytering till randomiserade studier är vanligtvis svårare än rekrytering till studier med annan design. Till den kvasiexperimentella studien söker forskaren ofta individer eller enheter som är intresserade av att prova den aktuella interventionen, varefter en jämförelsegrupp rekryteras. Till den randomiserade studien behöver forskaren nå människor i målgruppen som både är intresserade av att delta i interventionen och som kan tänka sig att bli randomiserade till en alternativ intervention, väntelista eller till en obehandlad kontrollgrupp. Det kan minska gruppen av potentiella studiedeltagare. Detta bekräftas av forskning som visar att just randomiseringen i sig kan vara ett hinder för medverkan – att man inte förstår bakgrunden till att den används, att man har en preferens för en viss intervention, känner osäkerhet inför en ännu inte utvärderad intervention eller att man inte vill riskera att hamna på väntelista eller i

kontrollgrupp (Ross, Grant, Counsell, Gillespie, Russell & Prescott, 1999).

> ***Ändrade förutsättningar ledde till låg statistisk power***
> Powerberäkningen för en extern verksamhetsutvärdering visade att 200 deltagare krävdes för att med 80 procents möjlighet hitta effekter av insatsen. Under studiens gång bestämmer styrelsen för den utvärderade verksamheten ett nytt och tidigare slutdatum för rekrytering som inte var förhandlingsbart. Trots ansträngningar lyckades forskarna endast rekrytera 128 deltagare. Analyser visar inte på några effekter av interventionen, men på grund av låg statistisk power går det inte att dra säker slutsats att interventionen saknar effekt. Forskaren har således genomfört en utvärdering som inte kan besvara forskningsfrågan (Bodin & Leifman, 2011).

En del av dessa hinder går att avhjälpa. Information om varför man randomiserar och vad det innebär behöver ges muntligt och skriftligt i samband med att man efterfrågar samtycke till medverkan (se avsnitt nedan). Hinder som har att göra med att personen vill ha en speciell intervention eller inte vill riskera hamna i jämförelse- eller kontrollgrupp är svårare att förebygga. En Cochrane-översikt om strategier för att förbättra rekryteringen till RCT som mestadels avsåg medicinska insatser visade att icke-blindade, öppna designer (där deltagarna vet vilken behandling de får) resulterar i högre andel rekryteringar än blindade, placebokontrollerade designer (där deltagarna inte känner till vilken behandling de får) (Treweek m.fl., 2010). Vid studier inom det beteendevetenskapliga och sociala området har forskaren oftast inte möjlighet att blinda deltagarna på grupptillhörighet. Baksidan med de öppna designerna är att de kan resultera i reaktioner från deltagarna som ger systematiska fel (eng. *bias*) i uppföljningsmätningarna, exempelvis som ett ökat bortfall i kontrollgruppen eller som socialt önskvärda svar från interventionsgruppen som vill gå forskaren till mötes för att man erhållit en resurs, oavsett vilken (jfr Hawthorne-effekten).

Om det är möjligt för klienter att få en insats utan att medverka i en studie så påverkar det valet att delta. Om insatsen redan är spridd

och forskaren når den presumtiva deltagaren efter att denne kommit i kontakt med vårdgivare, minskar rimligen intresset för att låta sig randomiseras. Om insatsen däremot är ny kan rekryterande verksamheter och möjliga deltagare informeras om att insatsen av etiska skäl måste beforskas, och därför endast ges inom ramen för studien. I den situationen är det dessutom ingenting som säger att insatsen som prövas är bättre än den som ges i jämförelsegruppen. Ibland kan dock rykten om en ny insats sprida sig i målgruppen. Detta kan underlätta rekryteringen men samtidigt öka risken för bortfall i kontrollgruppen, som kan känna sig missnöjd med att inte ha fått den nya interventionen. Deltagares preferenser är heller inte alltid som forskaren förväntar sig. I en studie av ett föräldraträningsprogram omfattande två versioner – en lång och en komprimerad – förekom att föräldrar sade sig föredra att hamna i den komprimerade insatsen på en dag framför den ordinarie med elva föräldraträffar.

Om man misstänker att rekryteringen kommer att bli svår och om det är etiskt försvarbart kan deltagarna i kontrollgruppen erbjudas någon form av ersättning. I två svenska studier av förebyggande skolprogram erhöll de skolor som randomiseras till kontrollgrupp (och därmed inte får den studerade insatsen) en summa pengar för sin medverkan. Skolorna fick själva välja vad pengarna skulle användas till så länge som det kom hela gruppen till godo, till exempel gå till klasskassan eller som ersättning till föreläsare för personalgruppen. Förutom att pengarna förefaller underlätta rekryteringsarbetet kan de även förebygga bortfall och systematiska fel av den typ som beskrivits ovan.

Andra egenskaper hos studien som man vet påverkar deltagandet är antalet möten och mättillfällen, besvärliga och omfattande mätprocedurer, förlorad arbetstid och resor till forskningscentra som kan innebära både tidsåtgång och extra kostnader. Man bör sträva efter att reducera praktiska besvär så långt det är möjligt och vid behov budgetera för att ersätta deltagarna för resekostnader och utebliven arbetstid. I USA har man exempelvis goda erfarenheter av

att erbjuda barnvakt och fria måltider för att rekrytera familjer till utvärderingar av föräldraträningsprogram.

Närståendes åsikter är ofta viktiga för en persons beslut att delta i forskning, och chanserna ökar när exempelvis en make, maka, partner eller vän ställer sig positiv till deltagande. En särskilt viktig roll i detta sammanhang har de professionella som involveras i en studie för att rekrytera deltagare genom sitt dagliga arbete. Patienter och klienter som litar på och har en bra relation till exempelvis sin läkare, psykolog eller kurator är mer benägna att delta i forskning än andra. Det finns också ofta en stor variation mellan professionella i en och samma studie om hur ofta man bjuder in patienter till att delta och hur många man rekryterar. Även om kunskapen om orsakerna till detta är bristfällig bör forskaren vara medveten om att många olika faktorer påverkar professionellas benägenhet att bjuda in och rekrytera patienter, även efter det att de accepterat medverkan i studien (Rendell, Merritt & Geddes, 2008). Det kan handla om att studieprotokollet avviker alltför mycket från sedvanlig behandling eller från den egna kliniska erfarenheten, eller att man upplever det som svårt eller tidsödande att informera patienterna om studiens design och innehåll. Professionella som inte hindras av sådana tvivel kan å andra sidan bli utmärkta ambassadörer för en studie.

Intervention

Erfarenheter visar att interventionens innehåll, format och omfattning påverkar rekryteringen. Dessa faktorer kan vara svåra att påverka inom ramen för en utvärdering där forskaren själv inte utvecklat interventionen, eftersom manualbaserade insatser bör utföras och utvärderas enligt sin manual. Anpassning av sådana egenskaper är mer aktuellt under utvecklingen av en intervention. Även efter sådana justeringar är det dock osannolikt att en insats är intressant för alla individer i målgruppen. En del föräldrar som tackade nej till att medverka i den finska utvärderingen av Beardslees familjeintervention uppgav att de inte ville prata med sina barn om den egna depressionen och därför inte ville delta (Solantaus, Paavonen,

Toikka & Punamäki, 2010). I en annan utvärdering som gällde behandling för alkoholberoende personer planerades 200 personer att randomiseras till antingen fem veckors slutenvårdsbehandling enligt tolvstegsmodell eller till sedvanlig öppenvårdsbehandling. Trots omfattande ansträngningar lyckades forskarna endast rekrytera 35 personer som var villiga att delta. Dessutom dök endast fyra av 18 patienter som randomiserats till slutenvårdsbetingelsen upp vid behandlingens början, medan bortfallet var litet i jämförelsegruppen. Det större bortfallet i slutenvårdsgruppen talar för att det var mindre attraktivt för målgruppen. Studien rapporterades som en pilotstudie och som ett exempel på en studie som misslyckades på grund av rekryteringsproblem.

Målgrupp

De individuella motiven för att delta i forskning varierar. Många har en altruistisk grundtanke; man förstår vikten av systematiskt kunskapssökande och vill genom sin medverkan lämna ett humanitärt bidrag. Vid studier av sekundärpreventiva och behandlande insatser finns ofta även ett personligt lidande hos målgruppen och man vill medverka i hopp om att få hjälp och lindring.

En systematisk översikt av faktorer som styr patienters val av medverkan i forskning visar att individfaktorer såsom sjukdomsgrad, utbildningsnivå och ålder inte påverkar benägenheten att delta i randomiserade studier på ett entydigt sätt. Dessa faktorers betydelse verkar snarare bestämmas av studiens egenskaper och sammanhanget i vilken den äger rum (Ross m.fl., 1999). Även om forskningen inte identifierar individfaktorer som entydiga bestämningsfaktorer för medverkan finns artiklar som vittnar om särskilda svårigheter vid rekrytering av utsatta målgrupper såsom åldrade sjuka, kulturella minoriteter, socialt utsatta eller personer med neurologiska skador (t.ex. Harris & Dyson, 2001; Bell, Hammond, Hart, Bickett, Temkin & Dikmen, 2008; UyBico, Pavel & Gross, 2008). I sådana fall blir det extra viktigt att forskargruppen och de som i praktiken tar de initiala kontakterna med möjliga deltagare har goda

kunskaper om målgruppen och agerar på ett och lyhört och förtroendeingivande sätt (Berger m.fl., 2009). Om forskaren inte har goda kunskaper om målgruppen är det viktigt att kontakta andra som har det för att ta del av deras erfarenheter. Individer kan även vara misstänksamma mot forskning och forskare på grund av egna eller andras dåliga erfarenheter. Ett ökänt exempel är Tuskegee-studien i USA som bidragit till misstänksamhet och ovilja till forskningsdeltagande bland afroamerikaner (Shavers, Lynch & Burmeister, 2002); 399 fattiga afroamerikaner med obehandlad syfilis följdes av forskare från 1932 till 1972, då projektet avslöjades av The New York Times. I mitten av 1940-talet blev penicillin standardterapi för syfilis, men patienterna informerades inte om denna botande behandling och gavs inte möjlighet att avbryta studien.

Ett vanligt hinder för medverkan är obehag eller rädsla inför att lämna ut personlig information som registreras. Försäkran om och efterlevande av forskningssekretess och konfidentialitet är basal och påverkar både sannolikheten att delta och validiteten i självrapportering. Om forskaren bedömer att konfidentialitet inte är tillräckligt och man vill ge deltagarna anonymitet, kan id-koder som genereras av deltagarna själva vara en möjlighet.

Information och samtycke

Både den procedur som ägnas åt att informera möjliga deltagare och informationens utformning påverkar en persons beslut om att delta eller ej. Om forskaren bemödar sig om att skapa förtroende och göra informationen förståelig för individer ur målgruppen, är det rimligt att detta både påverkar andelen som accepterar att medverka och andelen som stannar kvar i studien. Med utsatta och sköra målgrupper behöver man lägga sig extra vinn om proceduren för information och samtycke. I litteraturen rapporteras om goda erfarenheter av att avsätta ett informationsmöte med potentiella deltagare (Berger m.fl., 2009). Detta ger personer tid att ta del av informationen och att ställa frågor i sin egen takt. Forskaren ger även

tid till uppföljande frågor för att försäkra sig om att personen har förstått innebörden i informationen. Det är lämpligt att forskaren skapar en checklista över den information och de moment som behöver gås igenom under ett sådant möte. Detta för att försäkra sig om en hög kvalitet på samtyckesförfarandet och att alla personer får och förstår nödvändig information. Det finns studier som tyder på att proceduren för samtyckesförfarandet är ett förbättringsområde. En utvärdering av ett influensavaccin fann att endast 21 procent av deltagarna från socialt utsatta områden i Sydafrika kom ihåg att de randomiserats till olika behandlingar och att 81 procent inte hade förstått innebörden av placebobegreppet (Moodley, Pather & Myer, 2005).

Vad gäller utformning av information är det centralt att den är enkel, tydlig och korrekt samt anpassad till personens ålder och övriga förutsättningar. Om forskningen innefattar barn under 15 år krävs vårdnadshavares samtycke, medan barn över 15 år som inser vad forskningen innebär själva anses beslutskapabla. Man bör sträva efter att ge information såväl skriftligt som muntligt och att lämna tillräckligt med tid för funderingar och frågor (jfr kapitel 3).

Från beskrivningen av undersökningsproceduren ska personen kunna göra sig en realistisk bild av vad deltagandet kommer att innebära, både med avseende på de mätningar som ska göras och innehållet i de olika studiebetingelserna. Randomiseringsförfarandet kan exempelvis motiveras med att man fördelar deltagarna slumpmässigt på grupper, eftersom denna typ av studie ger säkrast information om en interventions effekter. En korrekt information utgör en säkerhet för den presumtive deltagaren som därmed vet vad han eller hon samtycker till.

Samtidigt som informationen ska ge en korrekt bild av vad det innebär att delta, är det viktigt att den också lyfter fram vad deltagare faktiskt kan vinna på att medverka, exempelvis att få möjlighet att prova en lovande metod. Det finns alltid en risk att sådan information drunknar i teknikaliteter som gör att deltagare avstår från att delta. Det är en balansakt att informera om möjliga förde-

lar som deltagande innebär utan att det slår över i marknadsföring. Förhoppningsvis har studien planerats så att alla har något att vinna på att delta.

Att deltagare har samtyckt efter att ha fått korrekt information underlättar även för forskare som personligen känner tveksamhet inför att randomisera deltagare till betingelser eller närma sig deltagare som är svåra att nå i samband med uppföljningar.

Det finns studier som tyder på att formen för samtyckesförfarandet påverkar rekryteringsgraden. I en australiensisk studie skickade man ut information om en planerad RCT gällande cancerscreening till patienter vid en primärvårdsmottagning. Ett passivt samtyckesförfarande resulterade i högre rekryteringsgrad (67 procent) än ett krav på aktivt samtycke (47 procent). I det första fallet kontaktade forskaren alla som inte avsagt sig intresse efter att ha fått informationen, medan man i det andra fallet endast kontaktade dem som själva meddelat ett intresse av att delta (Trevena, Irwig & Barratt, 2006). Det finns även studier där ett passivt samtycke inte bara innebär att man kan bli kontaktad och tillfrågad, utan att man faktiskt kommer med i studien om man inte aktivt säger ifrån. Exempel är utvärderingar av skolbaserade förebyggande insatser, där hela klasser eller skolor följs över tid, och där föräldrar informeras och ombeds säga ifrån via portofria svarskuvert om de inte vill att deras barn ska vara med. Det finns dock etiska problem med det passiva samtyckesförfarandet och ett aktivt samtycke krävs oftast för godkännande i forskningsetisk prövning.

Rekryteringsytor och kontaktsätt

Erfarenheter tyder på att forskaren bör behålla kontrollen över rekryteringen och delegera till professionella endast när detta är det enda alternativet. En orsak är att utvärderingen är primär för forskaren, men bara en av många prioriteter för den professionelle.

En annan orsak som tidigare berörts är risken för att professionella på förhand har en åsikt om vilken behandling en klient behö-

ver och därför inte erbjuder medverkan i utvärderingen. Professionella som av olika skäl upplever att rekryteringen är besvärlig kan bli bromsklossar i stället för ambassadörer för studien (Ross m.fl., 1999). Rekryteringsytor och kontaktsätt bestäms dock till stor del av studiens och målgruppens karaktär, och om en behandlingsmetod ska studeras i reguljär verksamhet involveras förmodligen behandlingsenheter och de professionella som arbetar där. Om en preventiv intervention ska studeras måste forskaren själv söka upp potentiella deltagare genom annonsering eller utskick. Oavsett sammanhang är det alltid viktigt att adressera den som har mandat att besluta om medverkan.

När det direkta rekryteringsarbetet ska bedrivas av professionella eller på flera platser är det klokt att budgetera för forskartid som innefattar koordination och kontinuerlig uppföljning av rekryteringsarbetet. Regelbundna kontakter med rekryterande professionella ger koordinatorn en bild av det totala och aktuella rekryteringsläget. Kanske går rekryteringen trögare på en enhet än på andra, och man kan behöva undersöka orsakerna till detta tillsammans med de berörda för att hitta lösningar. Ett exempel kommer från en svensk studie med låg rekryteringstakt i en av de medverkande kommunerna. I efterhand visade det sig att de involverade socialsekreterarna i den kommunen haft en handledare som ansåg att randomisering var oetiskt, och som därför avrått från att remittera ungdomar till studien. En mer kontinuerlig uppföljning hade gett forskarna en möjlighet att lyfta upp frågan till diskussion och kunde kanske ha ändrat socialsekreterarnas inställning. I andra fall kan det handla om att den reguljära verksamheten är så krävande för de professionella att rekryteringen kommer i skymundan. Forskaren behöver då på olika sätt hjälpa den rekryterande enheten. Kanske kan administrativt stöd i rekryteringsprocessen underlätta, till exempel i form av checklistor, där professionella kan bocka av olika moment som är nödvändiga i rekryteringsarbetet. Att regelbundet återkoppla jämförande information om rekryteringstakten till de ingående enheterna kan vara ett bra sätt att hålla liv i arbetet

samtidigt som forskarteamet får överblick över det totala läget. En kontinuerlig uppföljning av rekryteringsarbetet gör att svårigheter kan bearbetas snabbare och värdefull tid sparas in.

I de fall man rekryterar via skolor, socialkontor, vårdmottagningar och liknande bör forskaren vara medveten om att professionella ofta har en hög arbetsbelastning och egna agendor som konkurrerar med rekryteringen till studien. Erfarenheter visar att det är svårt för tjänstemän i sådana miljöer att vidmakthålla rekryteringsarbetet om inte deras chefer prioriterar det; vikten av att förankra studien hos enhetschefer kan inte nog betonas. Ibland kan forskaren vilja skriva en överenskommelse kring hur kontakterna mellan forskarteam, enhetschefer och tjänstemän ska se ut under studiens gång. Det är dock viktigt att inte nöja sig med en överenskommelse utan även kontinuerligt påminna chefer och tjänstemän om att studien pågår, exempelvis genom e-post, telefonkontakter samt personliga besök. Här behöver man vara beredd att avsätta mycket tid. Inom en studie där rekryteringen ägde rum inom socialtjänsten lade forskningsteamet mer än 100 timmar på olika former av möten med enheterna. Det kunde röra sig om korta samtal på 20 minuter upp till halvdagar med föreläsningar som behövdes för att upprätthålla intresset. Erfarenheter visar också att det är bra att ge rekryteringsarbetet en positiv inramning genom att exempelvis bjuda på kaffebröd eller annan uppmuntran då enheten lyckats rekrytera ett visst antal deltagare. En annan strategi är att ge enheter beting på hur många deltagare de ska rekrytera och när betinget uppnåtts behöver enheten inte rekrytera fler och kan tackas med någon form av belöning. Se dock till att från början vara tydlig med vad det är som belönas så att samarbetet kan flyta så friktionsfritt som möjligt – erhåller man exempelvis biocheckar efter rekrytering av 20 deltagare som samtyckt, eller efter rekrytering av 20 möjliga deltagare?

Den generella kunskapen om effekterna av olika rekryteringsstrategier vid randomiserade kontrollerade studier är begränsad. En systematisk översikt från Cochrane Collaboration identifierade visserligen 27 primärstudier, men få av dessa var så lika att man kun-

de väga samman effekter med metaanalys (Treweek m.fl., 2010). En norsk studie som ingick i översikten har undersökt effekten av telefonpåminnelser vid rekrytering av sjukskrivna till en intervention för återgång i arbetslivet. De ungefär 500 personer som inte svarat på en skriftlig studieinformation och inbjudan att delta (71 procent av de inbjudna) randomiserades till att följas upp med telefonpåminnelse eller ingen påminnelse; telefonpåminnelse resulterade i en nästan trefaldig ökning av antalet rekryterade (Nystuen & Hagen, 2004).

En annan Cochrane-översikt har vägt samman resultat från 513 studier av metoder avsedda att öka svarsfrekvensen i enkätundersökningar (Edwards, Roberts, Clarke, DiGuiseppi, Wentz, Kwan m.fl., 2009). Även om metoder för rekrytering till randomiserade studier och enkätundersökningar inte är direkt jämförbara, är resultaten relevanta när det är aktuellt att använda postutskick för att nå presumtiva deltagare. Majoriteten av studierna i översikten avser frågeformulär som sänds via post (n = 481) och resterande avser elektroniska formulär (n = 32). Översikten visar bland annat att svarsfrekvensen för postade frågeformulär ökar signifikant när man utformar materialet mer personligt, exempelvis genom att använda en handskriven namnunderskrift eller handskriven adressetikett, och frimärken snarare än förtryckt frankering för bifogade svarskuvert. Däremot fanns inga effekter av att använda färgat bläck, kolorerade papper eller illustrationer. Det verkar alltså inte finnas något att vinna på att göra materialet ”klatschigare”. Svarsfrekvensen ökar även när ett universitet snarare än en statlig eller kommersiell organisation står som huvudman för undersökningen och när deltagarna ersätts med pengar eller andra favörer.

Utformningen av annonser kan också påverka deltagandet. I en svensk studie av behandling för spelberoende märkte forskarna att rekryteringen underlättades när man formulerade om annonserna från ”Är du spelberoende och vill ha hjälp att lära dig hantera detta?” till ”Är du nyfiken på dina spelvanor?”. Att formulera sig på ett sätt som inte upplevs som stigmatiserande kan göra att fler ur

målgruppen söker sig till studien. Likaså kan man i annonseringen vända sig till anhöriga, släkt och vänner (”Känner du någon som spelar för mycket?”), då även anhöriga kan se annonsen och skicka vidare tipset om att anmäla sig. Forskare kan även ta hjälp av PR-byråer för att få hjälp med att upprätta kommunikationsplaner och strategier för annonsering.

I studier där rekryteringsprocessen helt eller delvis förläggs till internet finns möjlighet att både sänka kostnaderna och nå en större mängd individer. Annonser i dagspress kan exempelvis hänvisa till en hemsida med information om studien och preliminär screening av potentiella deltagare. För individen kan detta ge en ökad flexibilitet eftersom de olika stegen i rekryteringen kan skötas på tider som passar den enskilde. Dessutom kan kostnaderna sänkas när man screenar på basala urvalskriterier på en webbsida. Effektstudier som jämför olika kontaktsätt (t.ex. internet, annonsering i dagspress, personliga kontakter) är dock en bristvara, och rekrytering via internet har genererat såväl bra som dåliga resultat. Erfarenheter talar för att det är bra att utforma den webbaserade rekryteringsprocessen (sidor för studieinformation, screening etc.) i samråd med individer från den tänka målgruppen, att pilottesta den färdiga versionen för att ge en bild av rekryteringstakten och att länka till studien från väletablerade webbsidor som målgruppen kan tänkas besöka. Om rekryteringen via en webbsida är komplex och innefattar flera steg, är det bra med frekvent återkoppling till den sökande om var i processen han eller hon befinner sig.

Sammanfattning

Rekrytering av undersökningspersoner är en central del i en effektutvärdering och ofta förenad med svårigheter. Procedurer som kan underlätta rekryteringsarbetet är:

- Testa rekryteringsprocessen i förväg för att undersöka den förväntade inkluderingstakten och möjliga fallgropar i proceduren.
- Adressera den som har mandat att besluta om medverkan.

- Skaffa en bild av de organisationsfaktorer som kan hota rekryteringsprocessen och förebygg där det är möjligt (t.ex. via avtal).
- Behåll kontrollen över rekryteringen och delegera den till professionella endast när detta är det enda alternativet.
- Vid rekrytering via miljöer såsom skolor, socialkontor och vårdmottagningar, förankra med chefer och lägg mycket tid på vidmakthållande genom e-post, telefonsamtal, personliga besök och uppmuntrande belöningar. Studien som är viktig för forskaren behöver inte vara det för den enskilde tjänstemannen.
- Om forskare inte har goda kunskaper om målgruppen är det viktigt att kontakta andra som har det för att ta del av deras erfarenheter. Sök i databaser efter andra forskares publicerade erfarenheter av rekrytering av målgruppen.
- Utforma utskick via post eller e-post personligt och kontakta även dem som inte svarar initialt. Formulera annonser på ett ickestigmatiserande sätt och överväg att även rikta dem till anhöriga.
- Reducera de praktiska besvären för deltagarna och överväg olika former av ersättning.
- När professionella ansvarar för rekrytering bör forskaren fortlöpande följa detta för att tidigt kunna fånga upp svårigheter, ge stöd och eventuellt anpassa rekryteringsstrategin.
- Avsätt gott om tid för samtyckesförfarandet så att möjliga deltagare kan ta del av studieinformationen och ställa frågor i sin egen takt.
- Lyft fram vad deltagare har att vinna på att delta i studien, och ge målgruppsanpassad, tydlig och korrekt studieinformation.

Fördjupningslitteratur

Berger, L. K., Begun, A. L. & Otto-Salaj, L. L. (2009). Participant recruitment in intervention research: Scientific integrity and cost-effective strategies. *International Journal of Social Research Methodology, 12*, 79-92.

Grant, J.S., Raper, J.L., Kang, D-H. & Weaver, M.T. (2008). Research participant recruitment and retention. In A.M. Nezu & C.M. Nezu (Eds.), *Evidence-based outcome research. A practical guide to conducting randomized controlled trials for psychosocial interventions* (pp. 155-177). New York: Oxford University Press.

Referenser

Bell, K. R., Hammond, F., Hart, T., Bickett, A.K., Temkin, N. R. & Dikmen, S. (2008). Participant recruitment and retention in rehabilitation research. *American Journal of Physical Medicine & Rehabilitation, 87*, 330–338.

Berger, L. K., Begun, A. L. & Otto-Salaj, L. L. (2009). Participant recruitment in intervention research: Scientific integrity and cost-effective strategies. *International Journal of Social Research Methodology, 12*, 79–92.

Bodin, M. C. & Leifman, H. (2011). A randomized effectiveness trial of an adult-to-youth mentoring program in Sweden. *Addiction Research and Theory, 19,* 438–447.

Bodin, M. C. & Strandberg, A.K. (2011). The Örebro prevention programme revisited: A cluster-randomized effectiveness trial of programme effects on youth drinking. *Addiction, 106,* 2134–2143.

Edwards, P. J., Roberts, I., Clarke, M. J., DiGuiseppi, C., Wentz, R., Kwan, I., m.fl. (2009). Methods to increase response to postal and electronic questionnaires. *Cochrane Database of Systematic Reviews, 3*. Art. No.: MR000008. DOI: 10.1002/14651858.MR000008.pub4.

Harris, R. & Dyson, E. (2001). Recruitment of frail older people to research: lessons learnt through experience. *Journal of Advanced Nursing, 36*, 643–651.

Moodley, K., Pather, M. & Myer, L. (2005). Informed consent and participant perceptions of influenza vaccine trials in South Africa. *Journal of Medical Ethics, 31*, 727–732.

Nystuen, P. & Hagen, K. B. (2004). Telephone reminders are effective in recruiting nonresponding patients to randomized controlled trials. *Journal of Clinical Epidemiology, 57,* 773–776.

Ogden, T. & Halliday-Boykins, C. A. (2004). Multisystemic treatment of antisocial adolescents in Norway: Replication of clinical outcomes outside of the US. *Child and Adolescent Mental Health, 9*, 77–83.

Rendell, J. M., Merritt, R. K & Geddes, J. (2007). Incentives and disincentives

to participation by clinicians in randomised controlled trials. *Cochrane Database of Systematic Reviews*, *2*. Art. No.: MR000021. DOI: 10.1002/14651858.MR000021.pub3.

Ross, S., Grant, A., Counsell, C., Gillespie, W., Russell, I. & Prescott, R. (1999). Barriers to participation in randomized controlled trials: A systematic review. *Journal of Clinical Epidemiology, 52,* 1143–1156.

Shavers, V. L., Lynch, C. F. & Burmeister, L. F. (2002). Racial differences in factors that influence the willingness to participate in medical research studies. *Annals of Epidemiology, 12*, 248–256.

Solantaus, T., Paavonen, E. J., Toikka, S. & Punamäki, R-L. (2010). Preventive interventions in families with parental depression: Children's psychosocial symptoms and prosocial behaviour. *European Child and Adolescent Psychiatry, 19*, 883–892.

Sundell, K., Hansson, K., Andrée Löfholm, C., Olsson, T., Gustle, L-H. & Kadesjö, C. (2008). The transportability of multisystemic therapy to Sweden: Short-term results from a randomized trial of conduct-disordered youth. *Journal of Family Psychology, 22*, 550–560.

Trevena, L., Irwig, L. & Barratt, A. (2006). Impact of privacy legislation on the number and characteristics of people who are recruited for research: A randomized controlled trial. *Journal of Medical Ethics, 32,* 473–477.

Treweek, S., Pitkethly, M., Cook, J., Kjeldstrøm, M., Taskila, T., Johansen, M., m.fl. (2010). Strategies to improve recruitment to randomised controlled trials. *Cochrane database of systematic reviews*, *Issue 4.* Art. No.: MR000013. DOI: 10.1002/14651858.MR000013.pub5.

UyBico, S. J., Pavel, S. & Gross, C. P. (2007). Recruiting vulnerable populations into research: A systematic review of recruitment interventions. *Journal of General Internal Medicine, 22*, 852–863.

Kapitel 9 | Bjørn Arild Kristiansen, Christine Hassel Kristoffersen, Kristin Nordahl & Asgeir Røyrhus Olseth[1]

Praktiskt genomförande[2]

Randomiserat kontrollerade utvärderingar och andra effektutvärderingar använder en tydlig design med specifika krav på hur studien ska genomföras och hur resultaten ska rapporteras. En bra projektorganisation, noggranna förberedelser och ändamålsenliga rutiner för forskningsimplementering och hur utvärderingen genomförs har därför stor betydelse för både kvaliteten på data och resultatens validitet. Att säkerställa att data håller den föreskrivna kvaliteten är uppgifter som kräver god översikt och systematiska arbetsmetoder. Få har skrivit om denna typ av forskningslogistik. Utifrån egna erfarenheter har vi i detta kapitel systematiserat logistiken kring effektutvärderingar. Kapitlet behandlar områden som har särskild betydelse för det praktiska genomförandet av utvärderingar.

1 Logistikteamet vid Atferdssenterets forskningsavdelning har skrivit detta kapitel. De har tio års erfarenhet av att praktiskt planera, implementera och administrera data (data management) från 11 experimentella och longitudinella studier med population från 100 till 15 000 deltagare.

2 Kapitlet har översatts från norska till svenska av Malin Hultman, Socialstyrelsen.

Förberedelser

Projektorganisation

Grunden för att lyckas med att praktiskt genomföra en effektstudie är en tydlig projektorganisation (Stouthamer-Loeber & Van Kammen, 1995). Detta innebär först och främst att tydliggöra vem som har det överordnade ansvaret som projektledare, om det är en eller flera som ska ha forskningsansvar, vem som ansvarar för dataadministration och vem som har logistik-, personal- och budgetansvar. En projektplan beskriver vem som ansvarar för vad i utvärderingen. Centrala forskningsmedarbetare med kompetens inom logistik och dataadministration kan tillsammans med forskarna utgöra en projektgrupp som planerar arbetet och fattar viktiga beslut under arbetets gång. Redan från start kan projektgruppen säkerställa att praktiska och logistiska utmaningar hanteras, som rekrytering av deltagare, rutiner runt datainsamling och dataadministration. En bra investering, om de ekonomiska ramarna tillåter det, är att utse en projektkoordinator med ansvar för att ha uppsikt över och koordinera inblandade personer och uppdrag i projektets olika faser. En effektiv projektorganisation innefattar också genomtänkt planering av e-postrutiner och dokumentationssystem för att säkra ömsesidigt goda samarbets- och kommunikationsrutiner. Fasta mötestider för projektgruppen bidrar till att säkra informationsutbyte och nödvändig processutvärdering av kritiska element som exempelvis rekryteringsrutiner, datainsamling och att förebygga bortfall under arbetets gång.

Projektledning

En bra projektstyrning kräver en genomarbetad projektplan som ger information om och översikt över projektets bakgrund, mål och metod. Projektplanen utarbetas av den eller de ansvariga forskarna och utgör grunden för alla ansökningar om etiska tillstånd och finansiering.

Projektplanen kan också ha andra funktioner. *Internt* kan den vara

till praktisk hjälp i planeringsarbetet och fungera som ett överordnat styrdokument för genomförandet och för prioriteringar. Vidare kan projektplanen synliggöra de olika delarna i projektet och hur dessa relaterar till det övergripande målet. Den kan också ligga till grund för att revidera datainsamlingsmetoder och arbetssätt.

Externt kan en projektplan fungera som en informationskälla i kommunikation och i samarbete med andra forskare och organisationer som kommuner, skolor och behandlingsenheter. Projektplanen kan också användas för att informera om projektet till journalister och andra externa intresserade.

Utöver projektplanen behöver projektgruppen ett internt arbetsdokument som säkrar planering och styrning på en mer detaljerad nivå för alla faser i projektet. För att den ska vara användbar är det viktigt att den interna arbetsplanen används aktivt på projektgruppens möten och revideras vid behov. Rätt använd kommer arbetsplanen att fungera som dokumentation av det som planeras och genomförs.

En bra projektplan är övergripande och innehåller:

- Bakgrund till studien
  - Syfte och övergripande mål
  - Teoretisk bakgrund
  - Frågeställningar/hypoteser
- Metod
  - Population
  - Inklusions- och exklusionskriterier
  - Rekrytering
  - Information och samtycke
  - Randomiseringsförfarande
  - Kontrollgrupp
  - Informanter
  - Mätinstrument
  - Tillvägagångssätt för datainsamling
- Sekretess och forskningsetiska aspekter
- Tidsplan och finansiering
- Projektorganisation

Skydd av personuppgifter, data och dokumentation

Alla forskningsprojekt måste genomgå en forskningsetisk prövning som tillvaratar deltagarnas intressen. En ansökan måste göras i god tid före projektstart eftersom det kan krävas justeringar för att få ett godkännande. I studier där datainhämtningen är knuten till bestämda interventioner och tidpunkter, som exempelvis skolstart eller behandlingsstart, är det extra viktigt att en ansökan görs så tidigt som möjligt. Alla deltagare i studien måste skriva under ett informerat samtycke som är godkänt av etikprövningsnämnden. Samtycket beskriver vad deltagarna tillfrågas om att delta i, om det är observation, intervju eller ifyllande av frågeformulär. Samtycket ska också beskriva möjligheten att avsäga sig från deltagande i studien. Information om tidsperspektiv för deltagandet, hur data bevaras och när data ska raderas ska också ingå i samtyckesformuläret (se även kapitel 3).

En hög datasäkerhet krävs i alla forskningsprojekt. Skriftliga rutiner för hur det ska säkerställas måste finnas när det praktiska arbetet för att genomföra projektet planeras och för den forskningsetiska prövningen. Förteckningar över deltagarnas namn och adresser ska alltid bevaras skilda från alla forskningsdata. Både vid rekrytering av deltagare, under datainsamling och förvaring ska alla data vara säkrade och lagrade enligt projektansökan och det forskningsetiska godkännandet. Detta gäller med hänsyn till personskyddet för den enskilda deltagaren, men också för att det vid behov ska vara möjligt att återskapa databasen utifrån originaldokument. Alla offentliga dokument kring den etiska prövningen, inklusive projektplan och bilagor, annan korrespondens och beslut, lagras i ett säkert projektarkiv som är tillgängligt för alla projektmedarbetare. Systematisk arkivering av sådan dokumentation säkrar och förenklar rapportskrivandet och förenklar överföring av central information om några medarbetare slutar och nya ska anställas.

En intern arbetsplan är detaljerad och kan innehålla:

- Ansvarsområden och arbetsfördelning inom projektgruppen
- Mötesplan för projektgruppen
- Kommunikationsrutiner mellan projektets medarbetare
- Detaljerad tidsplan för att genomföra projektet
- Detaljerad budget
- Översikt över ansvar, arbetsuppgifter och rutiner på områden som:
  - Rekrytering
  - Dataadministration
  - Datainsamling och förebyggande av bortfall
  - Lagring och säkerställande av data
  - Uppföljning av anställda

Datasystem för översikt och kontroll

Under genomförandet av en effektutvärdering behövs ett system för att hålla uppsikt över alla deltagare och all datainsamling. För enklare studier kan ett vanligt Office-paket (t.ex. Excel, Access, FileMaker) vara tillräckligt, men för mer komplexa studier med många deltagare, flera grupper, olika mättillfällen och flera typer av informanter är det lämpligt med ett robust och skräddarsytt datasystem. Alla utvärderingar har sina speciella behov, så det kan vara svårt att hitta ett färdigt system som är tillräckligt bra. I de flesta fall måste sådana system utvecklas eller anpassas av personer med kompetens inom systemutveckling. Ett sådant system ska administrera och övervaka datainsamling, ge information om deltagare och respondenter och medverka till att den information man söker är lättillgänglig. Med ett databassystem kan också flera personer använda systemet samtidigt oavsett var de befinner sig. Detta kan vara till stor hjälp om man använder sig av datainsamlare som arbetar ute på fältet.

God kontroll över datainsamlingen är nödvändig i varje forskningsstudie och kanske speciellt i kontrollerade utvärderingar med en interventions- och en kontrollgrupp och där datainsamling ofta sker vid varierande tidpunkter. Det måste finnas en översikt över när och eventuellt hur datainsamlingen ska genomföras. När data-

insamlingen genomförts ska det registreras så att det i efterhand ska gå att se när data har samlats in och om frågeformulär och eventuella videoinspelningar kommit in. En fördel med ett sådant system är att tidpunkter för när data ska samlas in kan sättas upp så att datasystemet förvarnar när nästa datainsamlingstillfälle närmar sig. Detta är särskilt värdefullt när man har flera mättillfällen.

Alla som ska samla in data till projektet kan koppla sig direkt till systemet och få information om vilka datainsamlingar de ska göra. Därefter kan de själva rapportera tillbaka till systemet om vad som har genomförts. Datainsamlingar i systemet kan ha en statusvariabel som anger datainsamlingens framåtskridande, till exempel ”obehandlad”, ”avtalad”, ”genomförd”, ”gett upp” eller ”försöker upprätta kontakt” samt datum. Skickar man formulär per post kan andra statusmeddelanden vara aktuella, som ”skickat per post” och ”svar mottaget”. Systemet kan också rapportera om förseningar i datainsamlingen, så att den ansvarige datainsamlaren kan prioritera dessa. Ett alternativ är att den projektansvarige uppmärksammas och följer upp ärendet. Därmed ökar möjligheten att datainsamlingen genomförs när den är tänkt att ske.

I studier med hundratals deltagare är arbetet med att registrera och upprätthålla information om dem en betydande uppgift. Ofta vill man registrera kontaktinformation, randomiseringsgrupp och status för deltagande, till exempel ”deltar”, ”väntar på samtycke” och ”tackat nej”. Det är också värdefullt att kunna se en deltagares ”historia” med information om vilka datainsamlingar som har genomförts, ändringar i deltagares status och gärna öppna kommentarer. Datainsamlare ute i fält kan själva uppdatera informationen i databasen i stället för att kontakta en projektansvarig. Med färre led i informationsöverföringen blir informationen bättre uppdaterad och sparar tid för både datainsamlare och projektansvarig. Genom att lagra data centralt på ett ställe underlättas arbetet med att ha en uppdaterad överblick över datainsamlingen och informationen om det sker ändringar i studien. Det kan vara lämpligt att vissa funktioner endast kan ändras av den projektansvarige, som att ra-

dera deltagare från utvärderingen eller att ge upp försöket att följa upp en deltagare.

Rekrytering av deltagare

Rekryteringen av deltagare till studien beror på deltagargruppen och hur man planerar att nå den (se även kapitel 4). Rekryteringsarbetet måste förberedas utifrån vilka grupper man önskar att ha med i studien. Både frågor om hur man ska rekrytera och vem samt hur många som ska rekryteras avgör hur man planerar att genomföra rekryteringen. Informationen till potentiella deltagare måste anpassas och alla med rekryteringsansvar behöver utbildning som är anpassad för deras uppgifter. System för registrering av deltagare med kontaktinformation och id-nummer måste vara på plats innan rekryteringen startar. *CONSORT* (kapitel 18) har utarbetat ett flödesschema för rapportering av medverkan i randomiserade kontrollerade studier. De ställer krav på information om hur många deltagare som tillfrågats och hur många av dessa som är med i studien. Bland annat bör man registrera hur många som tackat nej, hur många som inte uppfyller inklusionskriterierna samt andra orsaker till att deltagare som blev rekryterade eller tillfrågade inte deltar i studien. Riktlinjer för registrering av information om dem som tillfrågades måste finnas tillgängliga och kända för dem som ska rekrytera deltagarna (Schulz, Altman & Moher, 2010; Moher, Hopewell, Schulz, Montori, Gotzsche, Devereaux m.fl., 2010).

Pilotstudie

En pilotstudie av viktiga element i utvärderingen kan vara en god investering under planeringsarbetet. När det gäller stora projekt kan man utvärdera till exempel rekryteringsrutiner, metoder för datainsamling, olika intervjuformulär, utbildningsmaterial, informationsbrev, val av kompensations- och belöningssystem eller annat som man vill veta hur det fungerar innan man startar själva studien. Anpassar man studien efter resultat i piloten får utvärderingen i allmänhet högre kvalitet.

Randomiseringsförfarande

I förberedelsearbetet för rekrytering av medverkande måste man i en randomiserad kontrollerad utvärdering också planera för att genomföra och dokumentera själva randomiseringen. Det finns olika sätt att randomisera och randomiseringen kan ske på olika nivåer. Randomiseringen kan ibland också misslyckas. Det finns exempel på att randomiseringen blir åsidosatt och att deltagare blivit placerade efter andra kriterier än slumpmässig fördelning eller att de som ska utföra den ändrar sina rutiner för att förenkla arbetet. Bra rutiner for randomisering och system minskar risken att randomiseringsförfarandet raseras (Boruch & Wothke, 1985; Shadish, Cook & Campbell, 2002). Det är viktigt att randomiseringen förankras hos alla medarbetare och samarbetande organisationer i projektet. I en del situationer och för personer som inte känner till forskningsmetodologin kring effektstudier kan randomisering uppfattas som orättvis. Det gör det extra viktigt att informera deltagare och dem som berörs av randomiseringen.

Exempel på randomisering

Ett exempel på ett randomiseringsförfarande som kan motverka komplikationer är det som användes i en RCT av interventionen *"Tidlig innsats for barn i risiko"*. I studien genomfördes randomiseringen genom att familjerna som rekryterades i kommunerna rapporterades till Atferdssenteret, där en anställd som inte deltog i rekryteringen utförde randomiseringen tillsammans med den huvudansvariga forskaren. Resultatet av randomiseringen förmedlades till behandlaren som tog kontakt med familjen. I denna studie blev randomiseringen genomförd efter att basmätningen var genomförd för att säkra att man hade basdata för alla som tackade ja till att delta (Kjøbli & Ogden, insänt för publicering). Det inträffar att deltagare som inte får de åtgärder de önskar avböjer medverkan i utvärderingen. Det är därför viktigt att basmätningen genomförts före randomiseringen så att man säkrar att information finns om alla. Om man dessutom genomför mätningen före randomiseringen minskar risken att deltagare låter sin besvikelse över vilken grupp de lottats till påverka svaren.

Id-koder och system för avidentifiering av data

I de flesta studier finns det krav på att data ska förvaras och analyseras så att ingen person är direkt identifierbar. Det betyder att man måste byta ut deltagarnas namn med en kod och att nyckeln till koden förvaras separat från data. Det finns många olika lösningar som kan användas, från den enklaste med löpande numrering av deltagare till system med mer komplexa koder.

En enkel lösning är att numrera deltagarna löpande. En annan möjlighet är att bygga koder genom att kombinera slumpmässigt valda bokstäver och tal (till exempel ”DL5289”). Med denna modell får man ett system där det finns många möjliga kombinationer, vilket gör att det är liten sannolikhet att fel angiven id-kod sammanfaller med id-koden för en annan medverkande i utvärderingen. Det gör det också möjligt att identifiera den rätta koden om det uppstår ett fel i delar av den.

Det är inte lämpligt att en kod innehåller beståndsdelar som representerar specifik information om respondenten, som till exempel region eller forskningsgrupp. Det kan tänkas att deltagare flyttar, eller att andra ändringar görs så att den ursprungliga koden blir missvisande. Ett annat problem med en sådan typ av id-koder är om personer kan identifieras utifrån information i koden.

Förberedelser för datainsamling

Arbetet med att samla in forskningsdata kräver noggranna förberedelser, detaljerade rutiner och funktionella system för datainsamling. Grundliga förberedelser kan bidra till att de ekonomiska resurserna utnyttjas bättre. Som en del av förberedelsearbetet av forskningsprojektet är det därför centralt att man undersöker alternativa lösningar och styrkor och svagheter med dessa. Bortfall av data kan få allvarliga konsekvenser för validiteten i en experimentell studie. Det gäller framför allt om bortfallet endast berör den ena undersökningsgruppen (Shadish m.fl., 2002). Primära delar i en utvärdering, som forskningsdesign, frågeställningar, grupper och

val av mätmetoder, styr det praktiska arbetet för datainsamlingen. Det kan till exempel vara strikta förfaranden knutna till hur ett datainsamlingsinstrument ska användas som begränsar möjligheten att använda elektroniska hjälpmedel eller bearbetning av data. Det finns också metoder för datainsamling som inte kan användas utan datainsamlare, som intervju och observation.

Kontakt med deltagare

Val av deltagare i en utvärdering har betydelse för hur kontakt ska tas och på vilket sätt data hämtas in. Till exempel kan språkproblem, en svår livssituation och geografiska avstånd vara faktorer som har betydelse för hur datainsamlingen genomförs. På vilket sätt man väljer att ha kontakt med deltagarna påverkar förberedelser och själva genomförandet. Ett val man måste göra är huruvida man ska ha personlig kontakt med deltagarna eller inte. Olika former av kontakt med deltagarna har styrkor och svagheter vad gäller att behålla deltagare i utvärderingen och kvaliteten på data man får in. Speciellt viktig är kontakten med deltagarna i en kontrollgrupp som inte får någon intervention, eftersom de utgör en särskild risk för bortfall.

Att använda datainsamlare i mötet med deltagarna ger bättre möjlighet att data faktiskt kommer in, att datainsamlingen sker vid rätt tidpunkt och att datamaterialet håller god kvalitet. Datainsamlare kan bland annat ge information direkt på plats. De kan också förhindra att frågor missförstås, eller att deltagare hoppar över eller missar frågor. Om deltagaren vill hoppa av projektet kan datainsamlaren fråga om orsaken och bidra till att förklara eventuella missförstånd. Att använda datainsamlare är emellertid ett resurskrävande sätt att samla in data på. Det kräver extra arbete i samband med rekrytering och upplärning, och det ökar utgifter till personalförvaltning och administration. Datainsamlare kan också till en viss grad påverka datainsamlingen om de uppträder på ett sätt som favoriserar en bestämd typ av intervention (Shadish m.fl., 2002). Metoder som att hålla datainsamlaren ”blind” för vilken grupp en deltagare tillhör kan bidra till att motverka en sådan situation, men det kan

vara svårt att genomföra i praktiken. När man hämtar in särskilt känslig information kan deltagare vägra att svara på frågor eller inte svara ärligt. Det gäller både vid intervjuer och vid användning av frågeformulär på stället.

Ett alternativ till att använda egna datainsamlare är att använda personer som redan är knutna till datainsamlingen i projektet, exempelvis terapeuter, lärare eller andra som är involverade i interventionen. Detta kan vara kostnadsbesparande för projektet och enklare för deltagarna, men det innebär utmaningar både när det gäller datakvalitet och praktiskt genomförande av datainsamlingen. För att en sådan lösning ska fungera måste man säkerställa att tillräckligt med tid är avsatt för datainsamling och att proceduren runt datainsamlingen och kraven på att insamlade data håller god kvalitet är explicita. Goda avtal med enheten där interventionen genomförs måste finnas för att säkra att datainsamlingen genomförs på rätt sätt. Utbildning och handledning ska göras på samma sätt som för projektanställda datainsamlare och kan bidra till att datainsamlingen fungerar i sådana situationer. Går det lång tid mellan datainsamlingarna är det speciellt viktigt med en användarvänlig manual för hur datainsamlingen ska genomföras. Om datainsamlaren också är deltagarens terapeut eller behandlare finns särskild risk för att detta kan påverka hur deltagaren svarar eller uppträder under datainsamlingen (Shadish m.fl., 2002).

Att samla in data utan personlig kontakt med deltagarna är mindre resurskrävande, vilket medför att man kan nå fler deltagare med mindre budget. Det vanligaste sättet att kontakta deltagarna utan att träffa dem personligen är med hjälp av frågeformulär. De kan skickas ut med vanlig post, via e-post eller delas ut av behandlare innan och efter interventionen. Utan den personliga kontakten med deltagarna ökar dock risken att formulär inte skickas tillbaka och bortfall av svar på enstaka frågor. Man har mindre möjlighet att följa upp deltagare för att kunna säkerställa svar på alla frågor och förklara frågorna om deltagarna inte förstår, och detta påverkar datakvaliteten.

Tekniska lösningar på datainsamling

Olika metoder för datainsamling innebär olika möjliga tekniska lösningar för de olika metoderna. Det vanligaste sättet att samla in data har hittills varit med hjälp av frågeformulär i pappersformat. Pappersformulär är ett enkelt verktyg som inte kräver särskild teknisk utrustning och kan därför användas av de flesta deltagare. Utformandet av frågeformulär kan göras på flera sätt, men det finns system för att behandla data i efterhand som kräver att frågeformulär är utformade på ett speciellt sätt.

Elektroniska lösningar för användning av frågeformulär har de senaste åren blivit mer stabila och tillgängliga. De elektroniska lösningarna har funktioner som kan göra ifyllandet av formuläret enklare och samtidigt bidra till att kvaliteten på data blir bättre. Man kan skräddarsy frågeformuläret efter deltagarnas svar, göra några frågor obligatoriska och ge möjlighet till att återuppta formuläret efter ett avbrott. En annan fördel med elektroniska frågeformulär är att data oftast kan exporteras direkt till format som kan läsas av de vanligaste statistikprogrammen. Detta sparar tid i bearbetningen av data. En utmaning vid sådana lösningar är att det kan hända att deltagare inte kan eller vill använda dator vid ifyllandet, och de kan vara skeptiska till att svara över internet med hänsyn till sekretess. Det kan också uppstå utmaningar avseende kontakten med deltagarna via till exempel e-post. Om man inte automatiskt får meddelande om att tekniska problem har medfört att frågeformuläret inte har nått fram till deltagaren, eller om en deltagare inte längre använder angiven e-post, kan värdefulla data gå förlorad.

Strukturerade intervjuer, eller assisterad ifyllning av frågeformulär, kan genomföras med hjälp av allt från frågeformulär i pappersformat eller elektroniskt till specialutvecklade programvaror för detta. Här finns det möjlighet att spara både pengar och tid genom att använda tekniska lösningar där data lagras i mer eller mindre analysfärdiga filer. När man har en forskningsmedarbetare på plats under ifyllandet, kan denna få specialutbildning i användning av elektroniska intervjulösningar eller internetbaserade formulär. På detta sätt

Checklista för elektroniska frågeformulär

- Tillräckligt flexibelt för att hantera olika standardiserade frågeformulär
- Användarvänlighet för deltagarna
- God säkerhet
- Möjlighet att återuppta avbrutna frågetillfällen
- Export av rådatafiler
- Möjlighet för obligatoriska frågor
- Möjlighet för olika vägar genom frågeformuläret
- Möjlighet att inhämta data när som helst
- Möjlighet att övervaka respondenterna (hur många svar, vem har svarat)
- Möjlighet för uppföljning av eller påminnelse till dem som inte har svarat
- Anpassat för de vanligaste webbläsarna

kan man undvika dubbelarbete under bearbetningen. Lösningar där varje deltagare kan ledas genom frågeformuläret på en dator, eller där intervjuaren lägger in deltagarens svar direkt i ett dataprogram (Computer-aided personal interview – CAPI), kan vara bra i sådana situationer. Speciallösningar för assisterat ifyllande av formulär i grupp (t.ex. elever i ett klassrum) kan vara lämpliga när det finns behov av att hjälpa dem som ska svara att förstå frågorna.

Observation som metod för datainsamling kräver direktkontakt med deltagarna. Hur observationsdata samlas in avgörs utifrån vilket beteende man önskar observera och på vilket sätt det observerade beteendet ska kodas. I huvudsak genomförs observationer på två olika sätt. Det första är att observatören finns på plats och skattar beteenden eller händelser direkt när de sker, något som är vanligt vid klassobservation eller på förskola. Då inhämtas inte direkt personidentifierade data och man behöver inte tillåtelse från föräldrar till en hel skolklass eller förskolegrupp.

Det andra sättet är att spela in skeenden med hjälp av bandspelare eller videokamera. Videoinspelningar är vanliga vid observation av exempelvis samspel mellan föräldrar och barn som utförs efter i förhand definierade uppgifter. Kodar man direkt på plats måste interbedömarreliabilitet kontrolleras genom att två observatörer vid

samma tillfälle kodar beteendet oberoende av varandra. När man använder videoinspelningar kan samma inspelning kodas av flera bedömare vid olika tidpunkter.

Om kodningen sker direkt i situationen, kan bärbara och lätthanterliga tekniska hjälpmedel användas. Kodning av videoobservationer kräver inköp av mer utrustning, både för inspelning, eventuell kopiering eller överföring av filmer och för själva kodningen. I det fallet kan det vara nödvändigt med speciell tillåtelse för digital lagring av film. Val av kodsystem styr rekrytering och utbildning av kodare. Vissa kodsystem kräver en bestämd typ av utbildning, och de flesta system kräver omfattande utbildning och träning. Antalet kodare, antalet observationer och längden på dessa avgörs av studiens omfattning, men också av tidsperspektivet och med hänsyn till reliabilitet mellan kodarna. Ett team med för få kodare ger svagare validitet, medan ett kodteam med för många medlemmar försvårar tillräckligt god interbedömarreliabilitet. Kodarna ska så långt som möjligt hållas blinda för deltagarnas grupptillhörighet och för om händelsen de kodar är genomförd före eller efter en eventuell intervention.[3]

Sätt att bearbeta data på

Manuell inmatning är den enklaste formen för inmatning av data och innebär att man läser av frågeformuläret och registrerar varje värde i en datafil eller ett dataprogram. Det finns olika tekniska lösningar för att förenkla processen och förbättra datakvaliteten. Fördelen med manuell inmatning är att den är enkel att genomföra och att det kräver relativt låg kompetens att använda de program som behövs. De största svagheterna med manuell inmatning är risken för att det kodas fel och att det tar lång tid att genomföra. Eftersom manuell inmatning av data har en hög felprocent behövs goda rutiner för att kontrollera datakvaliteten. Ett exempel på sådan kontroll är att data läggs in av två olika personer och därefter jämförs.

3 För mer information om observation, se Aspland & Gardner (2003), Nordal (2012) och Margolin, Oliver, Gordis, O'Hearn, Medina, Ghosh (1998).

Programvara som opererar med förhandsdefinierade värden på varje variabel kan få ner andelen fel, men det är trots det viktigt att kontrollera data för att undvika fel.

Optisk inläsning eller Optical Character Recognition (OCR) och Optical Mark Reading (OMR) är ett samlingsnamn för den teknologi som läser data optiskt från frågeformulär eller andra dokument. En typisk lösning för sådan elektronisk avläsning består av en scanner och programvara som överför data från papper till datafiler. Det finns också lösningar för att hämta in data direkt från fax eller scannade bilder som skickas elektroniskt. Optisk läsning kan drastiskt reducera tiden det tar att bearbeta data från frågeformulär till datafiler och det är mindre resurskrävande. Andelen fel blir också betydligt lägre än vid manuell inmatning. Det är dock skillnader i inläsningsprecision om pappersunderlaget handlar om rutor som kryssats i eller handskrivna bokstäver och siffror. En förutsättning för god användning av sådana lösningar är att man sätter sig in i de funktioner programmet man väljer har och skaffar kompetens eller hjälp så att det fungerar optimalt. En del system fungerar bäst när man har använt samma program under utformning och avläsning av frågeformuläret. Vanligtvis krävs inte särskild teknisk kompetens för själva inmatningen av data med optisk läsning. Men som regel är det nödvändigt att någon projektanställd övervakar processen och fattar beslut om oklarheter som rör till exempel kryss som hamnat utanför rutor, om flera rutor kryssats i och när det handlar om otydliga markeringar.

Optisk läsning reducerar antalet arbetstimmar för inmatning av data, men andra kostnader tillkommer i stället. Scanner och programvara kan vara dyrt, och det kostar att utbilda anställda eller att hyra in teknisk support. Högre kostnader för varje arbetstimme och större investeringar måste vägas mot fördelarna med att använda optisk läsning. Ett alternativ om man inte kan investera i teknisk utrustning kan vara att lägga ut bearbetning av data till ett företag som säljer sådana tjänster.

Direkt inmatning av data i elektroniska databaser är ett alternativ som blir allt vanligare. Det innebär att data blir tillgängligt i en

datafil samtidigt som det samlas in och där det inte behövs några särskilda aktiviteter för överföring eller inläsning. Det kan vara när deltagare fyller i ett frågeformulär på internet, eller då en forskningsmedarbetare matar in data under tiden en intervju pågår. Sådana lösningar kan innebära stora besparingar och möjliggör en bättre kvalitet på de data som kommer in eftersom det förhindrar att fel uppstår under överföring från papper till datafil. Det finns dock utmaningar som det är viktigt att vara medveten om. En är att det som vid optisk inläsning behövs speciell kompetens för att installera och använda dem. Det finns många leverantörer av internetformulär och intervjuplattformar eller program för standardiserade formulär, och det är viktigt att man sätter sig in i hur de fungerar. Kvaliteten i dem varierar, liksom priset. Eftersom det kommer att finnas behov av tekniskt stöd till dem som använder denna typ av lösning behöver resurser avsättas.

Forskningsimplementering

Lika viktigt som att implementera interventioner som ska utvärderas i studien är implementeringen av forskningsstudien. En lyckad forskningsimplementering i en randomiserad kontrollerad studie (RCT) kräver överblick och kontroll över kritiska faktorer i alla faser. Först och främst måste organisationen eller platsen där man ska implementera kartläggas så att faktorer som kan påverka forskningsarbetet och den praktiska genomföringen identifieras. Exempel på sådana faktorer är planerad omorganisation, ny ledarstruktur och juridiska reformer.

Förankringsarbete

När man har kartlagt organisationen där utvärderingen ska genomföras kan man starta det viktiga förankringsarbetet. Målet är att skapa förutsättningar för ett gott samarbete som genomsyrar alla nivåer i organisationen. Det är viktigt att projektet förankras på alla nivåer i organisationen så att alla får information om utvärderingens syfte

och genomförande. Ledningen behöver informeras så att den kan skapa förutsättningar för forskningsarbetet. Även andra bör informeras (som lärare, sjukvårdspersonal och behandlingspersonal), till exempel på informationsmöten med ledning och personalgruppen gemensamt. Informationsmöten där forskarna personligen presenterar projektet för alla intressenter främjar samarbete och en bra dialog. Forskarnas personliga närvaro är att föredra framför information via brev eller telefon. Bristande kunskap hos samarbetspartner kan leda till missförstånd och i värsta fall motstånd mot utvärderingsprojektet. Om motståndet dessutom kommer från personer som är informella ledare kan det få allvarliga konsekvenser.

Informationsmöten kan också vara viktiga för att identifiera nyckelpersoner i samarbetsorganisationerna (Blueprints for Violence Prevention, 2004). Dessa nyckelpersoner identifieras lätt genom sitt engagemang, sin roll och position och kan hjälpa till att marknadsföra projektet och bistå projektgruppen inåt i organisationen så att implementeringsarbetet och genomförandet av projektet underlättas. Dessa nyckelpersoner kan också bistå i både rekrytering och datainsamling och bör handplockas utifrån position, personliga egenskaper och motivation. Forskningsmedarbetare behöver anpassad utbildning och handledning i sin roll under hela projektperioden. Praktisk träning kan förbereda forskningsmedarbetare som ska träffa deltagare, antingen vid rekrytering eller vid datainsamling. Nyckelpersoner med en mer passiv roll är också viktiga eftersom de kan motivera andra att delta i utvärderingen. Det är viktigt att bygga goda relationer till nyckelpersonerna och upprätthålla dessa under projektets gång både genom skriftlig information och genom regelbundna informationsmöten. En god investering kan vara att kompensera forskningsmedarbetare som inte får lön genom projektet men som bidrar med sin insats utifrån sin anställning. Allt från böcker, choklad, nyhetsbrev om forskningsprojektet, lokala rapporter om resultat till inbjudningar till konferenser eller gemensamma kompetensdagar kan verka motiverande för att upprätthålla engagemanget.

Rekrytering och utbildning av forskningsmedarbetare

Rekrytering av kompetenta forskningsmedarbetare till olika uppgifter i projektet är lika viktigt vare sig det gäller heltidsanställda eller deltidsanställda. I en effektutvärdering varierar ofta behovet av arbetskraft. Finns det en bra arbetsplan kan man i olika faser av projektet utifrån den styra rekrytering, anställningar, utbildning och uppföljning. Olika arbetsuppgifter kräver olika kompetens, men med goda förberedelser kan man rekrytera personer som kan fylla flera roller under projektets gång. På en stor forskningsinstitution kan man också utnyttja stordriftsfördelar genom att medarbetare har samma typ av arbetsuppgifter för flera samtidigt pågående utvärderingar. Detta gäller särskilt för behandling av inkomna data och annan dataadministration.

Datainsamlare och andra som ska träffa deltagare personligen är projektets ansikte utåt och har därför en stor betydelse för deltagarnas intryck av forskningsarbetet. En relevant utbildning och tidigare erfarenhet av den aktuella målgruppen kan vara en fördel för datainsamlare. Andra viktiga egenskaper är noggrannhet, självständighet, flexibilitet och tålamod samt goda sociala färdigheter. Andra egenskaper kan behövas beroende på projekt. Ett bra träningsprogram för datainsamlare ger en helhetsförståelse av effektutvärderingen och datainsamlarens roll och arbetsuppgifter. Genom träning lärs färdigheter ut som behövs i arbetet på flera sätt, muntligt, skriftligt och gärna genom rollspel av aktuella situationer som datainsamlaren kan ställas inför. Vissa studier kräver att man behärskar filmutrustning, andra att man kan administrera urinprov. Säkerhetsrutiner för hembesök i utvärderingar där det finns risk för fysiskt våld eller liknande måste gås igenom. Systematisk träning baserad på en manual som regelbundet gås igenom och utvärderas säkerställer likvärdig kunskap åt alla vid alla tillfällen. Det är också viktigt med utbildning i databasarbete, om detta ingår i projektet. Detaljerade och användarvänliga manualer och ”kom ihåg-listor” till alla som ska använda databasen är en förutsättning för att den används korrekt. Utöver utbildningen är det viktig med återkommande möten

för datainsamlare där kvalitetssäkring av rutiner, vägledning och uppmuntran sätts i fokus, något som ökar motivationen och känslan av tillhörighet i projektet.

I vissa effektutvärderingar finns behov av att anställa forskningsassistenter för bestämda arbetsuppgifter som att förbereda och skicka ut frågeformulär, scanna och ”tvätta” data, koda observationer och liknande. Här är det viktigt att rekrytera rätt personer som passar till de arbetsuppgifter de ska utföra och som kan utföra arbetet när det finns behov, oavsett om det är kortare eller längre perioder. Utbildning av assistenterna varierar med arbetsuppgifterna, alltifrån några timmar för att scanna data till flera veckor eller månader om det handlar om att koda observationer. Oavsett arbetsuppgift måste det finnas skriftliga rutiner och en beskrivning av arbetet som ska utföras så att alla får samma information och att alla led i forskningsprocessen blir tillräckligt dokumenterade. För att kunna rekrytera och behålla bra forskningsassistenter är det viktigt med bra uppföljning så att assistenterna förstår att de gör ett viktigt arbete och tillhör ett team. Assistenterna är ofta deltidsanställda studenter som arbetar oregelbundna tider och som därför inte har möjlighet att bekanta sig med arbetsplatser och umgås med kollegor så som heltidsanställda projektmedarbetare kan. Att arrangera möten, temakvällar och gemensamma träffar med både ämnesrelaterat och socialt innehåll är därför god personalpolitik. Under en utvärderings gång kan det vara omsättning av forskningsmedarbetare, så det behövs goda rutiner som fångar upp att de som slutar ersätts med nya kvalificerade personer.

Exempel på vad en utbildningspärm för en datainsamlare kan innehålla:

1. Information om projektet
   – Studiens upplägg, roller och förfarande
2. Kontakt med deltagarna
   – Checklista för första mötet
   – Information och samtyckesblankett till deltagare
   – Rutiner for uppföljning av deltagare
3. Manualer for datainsamlare
   – Manual för eventuella tekniska hjälpmedel
   – Rutiner för intervjuer, videoinspelningar eller frågeformulär
   – Rutiner för att skicka, följa upp och påminna om frågeformulär

Daglig verksamhet

Datainsamling i praktiken

När alla planer är fastställda och rekrytering och träning av alla forskningsmedarbetare är genomförd kvarstår själva datainsamlingen. Gemensamt för alla är målet att skapa ett så komplett dataset som möjligt som utgångspunkt för fortsatta analyser.

En nyckeluppgift i denna viktiga fas i projektet är att de som ska samla in data följs upp så att man därmed förebygger att rutiner frångås. Med goda förberedelser, grundlig träning och konkreta beskrivningar av arbetsuppgifter och rutiner är man en bra bit på väg. Alternativa lösningar och krisplaner för *när* något går fel (vilket alltid sker) är nödvändigt i alla studier. Innan datainsamlingen påbörjas bör man ha prövat hur lång tid det tar och om det finns förhållanden som kan medföra att man måste lägga in extra tidsmarginaler. Detta är viktig information som måste förmedlas till dem som ska stå för själva datainsamlingen. Vid vissa tillfällen, som vid skolstudier, kan det vara förnuftigt att man har flera dagar avsatta till datainsamling, eftersom elever kan vara borta från skolan den dagen datainsamlingen är planerad.

När datainsamling genomförs är det viktigt att ha en loggbok eller liknande system för eventuella avvikelser som uppstår. Det kan

till exempel vara att datainsamlingen avbryts under pågående insamling eller att en observation inte kan genomföras på tänkt sätt. Om man observerar elever i klassrummet och många av eleverna inte är på plats måste det finnas möjlighet att rapportera detta. Det kan vara lämpligt att tänka igenom vilka avvikelser som kan förekomma och göra ett system för rapportering av dem.

Utöver loggbok och rapportering är det viktigt att ha färdiga rutiner för hantering av avvikelser. Vid avbruten datainsamling kan deltagaren få ett pappersformulär och ett färdigfrankerat kuvert med sig hem för att fylla i det som saknas. Det behövs också ett system som säkerställer att deltagaren påminns om inte svaren har kommit in inom rimlig tid.

En forskningsmedarbetare bör ha ett överordnat ansvar för att datainsamlingen går enligt plan och för att vidta åtgärder när det behövs. I en liten studie kan det vara forskaren själv, men i större studier kan uppgiften kräva en person på heltid. I studier med egna datainsamlare är handledning och uppföljning av dessa en betydelsefull arbetsuppgift som måste tas med i planeringen av projektet. I alla studier uppstår situationer som kräver förtydliganden och uppföljning och där oförutsedda händelser behöver hanteras. Information och kommunikation under projektets gång är viktig för att hitta bra lösningar när problem uppstår (se avsnittet om projektorganisation).

Att behålla deltagarna

Randomiserade kontrollerade utvärderingar varierar i omfattning när det gäller deltagargrupper, antal deltagare, varaktighet, antal datainsamlingstillfällen och tid mellan dem. Vissa studier har bara två insamlingstillfällen med en intervention av kort varaktighet mellan, medan andra följer grupperna över flera år för att undersöka långtidseffekter av en intervention. I sådana longitudinella utvärderingar kan det gå kort eller lång tid mellan datainsamlingstillfällena, men oavsett det är det en utmaning att behålla deltagarna i studien.

För att kunna följa upp deltagarna är det först och främst viktigt

med en översikt över vilka deltagarna är, samt var och hur de kan kontaktas. Kontaktinformationen måste uppdateras regelbundet så att inte deltagare försvinner på grund av byte av jobb, bostad eller telefonnummer. Upplysningar som kan vara bra att registrera till nästa datainsamlingstillfälle är tidpunkt för telefonkontakt, vilket telefonnummer som används, om deltagaren kommer att vara bortrest över en längre period och liknande. Om studien har en kontrollgrupp där deltagarna inte får någon intervention är det viktigt att vara extra noga med regelbunden uppdatering av kontaktinformation så att det går att nå dem vid nästa datainsamling.

Personlig uppföljning med samma intervjuare vid varje datainsamling är att föredra eftersom det kan göra att deltagarna känner sig väl omhändertagna, vilket motverkar bortfall. Detsamma kan positivt formulerade påminnelser inför datainsamling göra, antingen via telefon, sms, e-post eller via andra passande kanaler. Att datainsamlarna svarar på frågor om studien och förklarar eventuella missuppfattningar kan förebygga att deltagarna avbryter sin medverkan. Under datainsamlingen är det viktigt att få deltagarna att känna sig trygga i situationen så att de inte upplever datainsamlingen som obehaglig. Detta kan vara en utmaning vid exempelvis videoinspelning eller inhämtning av urinprov. Vissa gånger kan det vara på sin plats att bjuda på mat och dryck under en intervju. Andra gånger kan det vara mer passande att ordna transport till lokalen där insamling av data sker, ordna barnvakt eller ge ersättning för utlagd parkeringsavgift.

Att kompensera deltagarna för deras tid och ansträngningar kan öka motivationen till fortsatt deltagande. Kompensationen måste anpassas till deltagargruppen så att de upplever att de verkligen får en belöning. Ungdomar kan exempelvis ha andra preferenser än pensionärer eller kroniskt sjuka. Andra exempel på kompensation är kontanter, en bok, en cd-skiva, trisslotter, ett presentkort eller choklad. Den planerade kompensationen måste ingå i den etiska prövningen av projektet. I utvärderingar som sträcker sig över längre tid kan det löna sig att göra nyhetsbrev till deltagarna med innehåll som ger information om studien, nästa datainsamlingstillfälle och

liknande som kan vara motiverande för fortsatt deltagande. För att behålla deltagarna kan man också ge information genom en egen webbsida. Ett årligt lotteri bland alla deltagare, eller utskick av födelsedags- eller julkort, är andra idéer för att behålla deltagarna.

Utöver kompensation kan det vid vissa tillfällen vara viktigt att ge deltagarna eller deltagande organisationer feedback om utvärderingen efter att datainsamlingen är utförd. Det kan bidra till att deltagarna upplever värdet av att medverka i forskning.

Flexibilitet när det gäller både tidpunkter och plats för datainsamling kan göra det enklare för deltagare att medverka i studien. Om det passar bättre att få besök i hemmet för att svara på frågor kan det ge deltagarna en möjlighet att få tiden att gå ihop. Att erbjuda alternativa tekniska lösningar för att svara på ett frågeformulär, som e-post eller en postad pappersversion, kan också öka sannolikheten för att datainsamlingen äger rum.

Dataadministration

En rad aktiviteter kan göras för att säkerställa kvaliteten vid datainsamling. Detta kallas för dataadministration (eng. *data management*) och omfattar allt från att följa upp data från det att data samlas in till att göra ett dataset klart för analyser. Det inträffar alltid situationer under processen där fel eller oregelbundenheter upptäcks i datasetet. En del av dataadministrationen är att försöka att hantera detta och utveckla rutiner för kontroll och korrigering av sådana fel. Detta är en viktig uppgift som kan vara tidskrävande, men kvaliteten i datasetet stärks och man undviker frustration när data ska analyseras eller att felaktigheter påverkar resultaten.

När data samlas in har man redan en rad potentiella felkällor. Det kan handla om allt från ogiltiga svar i frågeformuläret till felaktiga id-koder. Val av frågeformulär och insamlingsmetoder har betydelse för vilka problem som kan uppstå. Ett vanligt problem är att respondenter sätter kryss mellan två svarsalternativ eller anger flera svarsalternativ där det bara ska vara ett. Med en datainsamlare närvarande kan det undvikas. Använder man elektroniska frå-

geformulär kan respondenten inte göra ett sådant fel. När man använder pappersformulär som deltagaren fyller i utan uppföljning måste man vara förberedd på att gå igenom varje formulär och leta efter oregelbundenheter och hantera dessa efter bestämda rutiner som utvecklas tillsammans med forskare eftersom analyserna kan påverkas av de val som görs. Flera rutiner utvecklas ofta under tiden som formulären kommer in, eftersom det kan vara svårt att förutse alla problem som kan uppstå. Alla rutiner bör dokumenteras, både så att man är konsekvent när man hanterar frågeformulären och så att alla kan se vilka metoder som har använts.

Det är lämpligt att konstruera en eller flera datafiler (*”masterfiler”*) som ger en helhetlig översikt av urvalet i studien. Denna översiktsfil kan innehålla viktiga variabler som ålder, kön, randomiseringsgrupp och så vidare. Översiktsfilen bör vara i samma format som de andra datafilerna, så att man enkelt kan lägga samman dessa. Översiktsfilen används vid felkontroll av datasetet, och kan användas av forskaren till att rapportera om urvalet.

När data är överförda från frågeformulären eller från observationer till datafiler, kan man börja kontrollera för eventuella fel. Denna process kallas att tvätta data. Innan man tvättar data måste man skaffa sig god kännedom om hela datasetet så att man kan identifiera kontrollpunkter och test som behöver utföras. Dessa kontroller varierar utifrån hur studien är upplagd och utifrån hur frågeformulären ser ut. Man vill försäkra sig om att nyckelinformation som id-koder, kön och ålder stämmer, att man har det korrekta antalet individer och att det inte har uppstått överföringsfel om man har stansat eller scannat frågeformulären. Om deltagarna exempelvis har fått ange årsinkomst, så bör man kontrollera att det är ett rimligt värde.

Datasetet organiseras i allmänhet med individer på rader och forskningsvariabler i kolumner. Varje rad i datasetet ska ha en eller flera variabler som ger en unik identifiering av den enskilda individen. Genom att leta efter dubbletter på dessa variabler kan man identifiera felaktiga id-koder eller fall som har blivit registrerade flera gånger. Om data kommer från scannade eller stansade pap-

persformulär bör man leta efter extremvärden (eng. *outliers*) i datasetet och kontrollera att dessa stämmer med pappersformuläret (se även kapitel 10). I elektroniska formulär har man ofta möjligheten att begränsa vilket värde respondenten kan skatta, och kan hindra att sådana fel uppstår från början.

När datasetet är komplett kan man slå samman olika datafiler (t.ex. olika frågeformulär eller olika datainsamlingstillfällen). Det är ett bra sätt att identifiera data som har fått fel id-kod under processens gång. Man kan kontrollera kön, ålder och liknande för att verifiera att det rör sig om samma individ.

Det lönar sig att ha ett bra, standardiserat system för att behandla datafiler. Rådatafilen bör inte ändras. Det är bättre att arbeta med kopior av filen. Skulle något fel göras under datatvätten, exempelvis att data felaktigt tas bort, ska det finnas en möjlighet att gå tillbaka till originalfilen. Det är också viktigt att ha ett system för att dokumentera vad som har gjorts under datatvätten. Alla ändringar som görs med filen ska registreras i detta system så att man kan rekonstruera vad som gjorts. De flesta statistikprogram som SPSS eller SAS har ett eget scriptspråk som rekommenderas att använda eftersom det i sig kan vara tillräcklig dokumentation. När datasetet är tvättat är det klart för analys, men i många fall strukturerat på ett annat sätt än det som forskaren önskar. Beroende på vilka analyser som ska göras måste filen struktureras på ett sätt som passar till analyserna. Det kan röra sig om att slå samman eller omorganisera filer, ändra skalor och invertera enskilda variabler. I många studier är det ett omfattande arbete som sker i nära samarbete med forskaren som ska utföra analyserna. Den sista punkten för det praktiska genomförandet av studien är som regel att hålla uppsikt över när data ska raderas eller förstöras och ansvara för att detta görs inom tidsfristen.

Avslutande kommentarer

I det här kapitlet har vi visat på komplexiteten i det praktiska arbetet med att genomföra en experimentell utvärdering. I mindre stu-

dier med få deltagare kan en ensam forskare hålla uppsikt över och kontrollera processen, men i de flesta studier finns ett behov av praktiskt stöd i större eller mindre omfattning. En bra planering, en god struktur av organisation och arbetsuppgifter, men även flexibilitet inom ramarna under arbetes gång är en bra modell för den här typen av praktisk forskningslogistik.

Några tips för att öka sannolikheten för en väl genomförd studie är:

- tydlig och genomtänkt projektorganisation
- bra planeringsarbete och pilotstudie
- systematiskt kommunikations- och förankringsarbete
- grundlig utbildning och uppföljning av datainsamlare
- välutprövade tekniska hjälpmedel och reservplaner
- detaljerade beskrivningar av datainsamling och rutiner kring denna
- rutiner och system för att kvalitetssäkra datainsamling och kontrollera datamaterialet.

Fördjupningslitteratur

Prinz, R. J., Smith, E. P., Dumas, J. E., Laughlin, J. E., White, D. W. & Barrón, R. (2001). Recruitment and retention of participants in prevention trials involving family-based interventions. *American Journal of Preventive Medicine 20.1, Supplement 1*, 31-37.

Stouthamer-Loeber, M. & Van Kammen, W. B. (1995). *Data collection and management a practical guide.* Thousand Oaks: Sage Publications.

Referenser

Aspland, H. & Gardner, F. (2003). Observational Measures of Parent-Child Interaction: An Introductory Review. *Child and Adolescent Mental Health*, 136–143.

Blueprints for Violence Prevention, Center for the Study and Prevention of Violence, University of Colorado-Boulder, 2004, Office of Juventile and Delinquency Prevention.

Boruch, R. F. & Wothke, W. (1985). Seven kinds of randomization plans for designing field experiments. *New Directions for Program Evaluation, 1985*, 95–113.

Kjøbli, J. & Ogden, T. A (under tryckning). A randomized effectiveness trial of Brief Parent Training in primary care settings. *Prevention Science*.

Margolin, G., Oliver, P. H., Gordis, E. B., O'Hearn, H. G., Medina, A. M., Ghosh, C. M., m.fl. (1998). The nuts and bolts of behavioral observation of marital and family interaction. *Clinical Child and Family Psychology Review*, 195–213.

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gotzsche, P. C., Devereaux, P. J., m.fl. (2010). CONSORT 2010 Explanation and Elaboration: Updated guidelines for reporting parallel group randomised trials. *J.Clin.Epidemiol*., 63, 1–37.

Nordahl, K. B. (2012). Systematisk observasjon av barns samhandling med andre. I Backe-Hansen & Frønes (red.) *Metoder og perspektiver i barne- og ungdomsforskning*, (s. 158–173). Oslo: Gyldendal Akademisk.

Schulz, K. F., Altman, D. G. & Moher, D. (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMC.Med.*, *8*, 18.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin Company.

Stouthamer-Loeber, M. & Van Kammen, W. B. (1995). *Data collection and management – a practical guide*. Thousand Oaks: Sage Publications.

Kapitel 10 | Martin Forster & Terje Ogden

Behandlingstrohet

I experimentella utvärderingar manipuleras den oberoende variabeln (t.ex. behandling jämfört med kontroll) i syfte att undersöka dess effekt på beroendevariabeln (t.ex. psykisk hälsa). Traditionellt har man i samhällsvetenskaplig forskning lagt stor vikt vid att ha god kontroll över beroendevariabeln, till exempel genom att säkerställa reliabilitet och validitet i utfallsmått. Däremot har anmärkningsvärt lite uppmärksamhet ägnats åt kontroll av den oberoende variabeln, det vill säga att inkludera tillförlitliga mätningar av behandlingstrohet (eng. *treatment integrity, treatment fidelity*). Det konstaterades exempelvis i en genomgång av alla interventionsstudier från sex vetenskapliga tidskrifter mellan år 2000 och år 2004 (Perepletchikova, Treat & Kazdin, 2007). Endast i 3,5 procent av totalt 147 artiklar var mätning och/eller redovisning av behandlingstroheten adekvat, trots att de sex tidskrifterna räknas som ledande inom sina fält. En annan analys av användandet av behandlingstrohet har kommit fram till liknande resultat (Naleppa & Cagle, 2010).

Det här kapitlet inleder med att diskutera argument för att mäta behandlingstrohet samt olika definitioner av begreppet. Därefter ägnas huvuddelen av kapitlet åt följsamhet, som är den mest grundläggande dimensionen av behandlingstrohet. Det är en stegvis genomgång som konkret visar hur man går tillväga för att mäta följsamhet. Många av principerna i den genomgången är generella och

gäller även vid mätning av andra dimensioner av behandlingstrohet. Därefter diskuteras mätning av behandlarnas kompetens. Det följs av en genomgång av mätinstrument för behandlingstrohet som har dokumenterade psykometriska egenskaper. Kapitlet avslutas med en diskussion om sambandet mellan behandlingstrohet och utfallsvariabler i effektstudier.

Varför är det viktigt att mäta behandlingstrohet?

Det primära skälet till att mäta behandlingstrohet är att säkra slutsatser från experiment endast kan dras om man har god kontroll över både oberoende och beroende variabler. Det stärker den interna validiteten genom att inflytandet av ovidkommande variabler minimeras. Även extern validitet påverkas. Låt oss ta ett konkret exempel. I Sverige planerades nyligen en effektutvärdering av *Aggression Replacement Training (ART),* som riktar sig till ungdomar med utagerande problematik. Det är en av de mest spridda interventionerna för den målgruppen i Sverige. I en inledande kartläggning visade det sig emellertid att 90 procent av ART-verksamheterna inte mötte minimikraven på behandlingstrohet för programmet (Kaunitz & Strandberg, 2009). En effektstudie av ART i dessa verksamheter skulle bli svårtolkad. Oavsett vilket resultat studien skulle visa så skulle det aldrig gå att hänföra till ART, eftersom 90 procent av deltagarna i studien i praktiken jobbar på något annat sätt (d.v.s. svag intern validitet). Om man trots det skulle sprida metoden, så skulle det vara omöjligt att förmedla innehållet i insatserna till andra vårdgivare. I den meningen blir även den externa validiteten eller generaliserbarheten svag.

Definition och dimensioner av behandlingstrohet

Det förekommer flera definitioner av behandlingstrohet, men begreppet brukar åtminstone omfatta två dimensioner: följsamhet (eng. *adherence*) och kompetens (eng. *competence*) (t.ex. Waltz, Ad-

dis, Koerner & Jacobson, 1993). Hög följsamhet innebär att specificerade moment i en behandling faktiskt har genomförts, till exempel enligt ett behandlingsprotokoll eller en manual. Hög kompetens innebär att dessa moment har genomförts på ett skickligt sätt, till exempel med rätt timing och flexibilitet. Följsamhet brukar beskrivas som kontextoberoende – det vill säga behandlingsmoment som är oberoende av vem som är klient eller under vilka omständigheter behandlingen sker (Barber, Sharpless, Klostermann & McCarthy, 2007). Kompetens handlar däremot bland annat om de kontextuella anpassningar som krävs när behandlingen tillämpas i enskilda fall. Följsamhet och kompetens syftar på behandlarens sätt att genomföra en intervention. Ibland omfattar mätningar av behandlingstrohet även hur klienten förstår och tillämpar behandlingen (Schulte, Easton & Parker, 2009). I tabell 10:1 beskrivs dimensioner av behandlingstrohet som förekommer i litteraturen. I detta kapitel kommer framför allt olika aspekter av följsamhet att beskrivas, eftersom det är en förutsättning för övriga dimensioner. Även kompetens kommer att diskuteras i ett separat avsnitt.

Mätning av följsamhet

I det här avsnittet beskrivs processen för att utveckla instrument för att mäta följsamhet. *Komet för föräldrar,* som är en intervention för föräldrar med utagerande barn, kommer att användas som genomgående exempel. Processen beskrivs steg för steg, men därmed inte sagt att alla steg alltid är nödvändiga eller att de måste ske i nämnd ordning. Även om utgångspunkten är att mäta följsamhet, så gäller principerna i stort också för andra dimensioner av behandlingstrohet (t.ex. kompetens).

Steg 1. Optimera följsamheten – eller bara mäta?

En första viktig fråga som rör följsamhet är vilken slags studie som ska genomföras. I en modellutvärdering (eng. *efficacy*), där syftet är att undersöka hur interventionen fungerar under optimala förhål-

Tabell 10:1. Vanliga dimensioner av behandlingstrohet.

Dimension	Definition	Exempel på mätning
Följsamhet (eng. *adherence*)	Andel centrala behandlingsmoment (kärnkomponenter) som genomförs.	En enkät med frågor som rör genomförande av kärnkomponenter.
Kompetens (eng. *competence* eller ibland *quality*)	Den skicklighet med vilken den specificerade behandlingen genomförs.	En expertskattning av kvaliteten på behandlingens genomförande, exempelvis hur väl en behandlare anpassar sättet att genomföra behandlingen efter klientens unika förutsättningar.
Differentiering (eng. *differentiation*)	En variant av följsamhet där även förbjudna/avrådda behandlingsmoment specificeras. Används vid studier som jämför olika behandlingar.	En expertskattning av huruvida behandling A innehåller kognitiva strategier (utan att innehålla psykodynamiska moment) och vice versa för behandling B.
Exponering (eng. *exposure*)	Antal, längd (tim/min) och frekvens med vilken sessionerna i en behandling genomförs. Kan betraktas som del av dimensionen följsamhet.	Journalanteckningar där uppgifter om sessioner bokförs. Eller insamling av uppgifter som del av en självskattning.
Klientförståelse (eng. *participant comprehension*)	Del av dimensionen följsamhet. Hur väl klienten förstår avgörande inslag i en behandling.	Test/enkät som mäter klientens kunskap.
Klientföljsamhet (eng. *participant responsiveness*)	Del av dimensionen följsamhet. Hur väl en klient deltar i och tillämpar behandlingen.	Skattning av klienters engagemang i en behandling eller andel genomförda hemuppgifter.
Strukturella faktorer	Utbildningsnivå hos personal etc.	Kvantifierade eller deskriptiva data som rör behandlarnas utbildning, erfarenhet och förutsättningar för att genomföra interventionen.

landen, kan och bör åtgärder vidtas för att säkra god följsamhet. Det kan ske genom utbildning och handledning av behandlarna. Följsamheten måste också mätas, helst före fördelning av deltagare till jämförelsegrupperna, för att säkerställa att behandlare som inkluderas i studien håller måttet. I en verksamhetsbaserad utvärdering (eng. *effectiveness*), där syftet är att undersöka hur interventionen

fungerar i reguljär verksamhet, får inga särskilda åtgärder vidtas för att öka följsamheten. I bästa fall lyckas behandlarna genomföra interventionen med tillfredsställande följsamhet som inte varierar för starkt. Det innebär en högre statistisk power och att resultaten blir möjliga att tolka. Om följsamheten däremot visar sig vara låg är det svårare att både finna och tolka eventuella signifikanta skillnader – som i det tidigare exemplet om ART. I en nyligen publicerad verksamhetsutvärdering av Komet prövades behandlingen i socialtjänsten, med den reguljära personalen som behandlare (Kling, Forster, Sundell & Melin, 2010). Personalen fick utbildning och handledning i den nya metoden enligt den modell som vanligtvis användes inom organisationen, men inget extra stöd därutöver. Behandlingen ägde rum inom ramen för personalens anställningar och i deras egna lokaler. Följsamheten mättes genom att behandlarna skattade hur stor del av manualen som de hade genomfört efter varje session. Föräldrarna som deltog fick även telefonsamtal eller fick fylla i en webbenkät efter varje session som kontroll av gruppledarnas uppgifter. Enligt gruppledarna genomfördes i snitt 76 procent av manualen, vilket var relativt nära föräldrarnas skattning (70 procent). Därmed bedömdes följsamheten vara tillräckligt god för att kunna dra säkra slutsatser av resultaten i studien.

Steg 2. Jämförs olika slags behandlingar?

Om en jämförande studie med flera aktiva behandlingar ska genomföras så kan det vara aktuellt att utgå från en utvidgad definition av följsamhet, nämligen differentiering (eng. *treatment differentiation*). Differentiering är i praktiken samma sak som följsamhet, men förutom att bestämma och mäta vad som ingår i en behandling (kärnkomponenter) så måste man också bestämma och mäta sådant som inte ingår i behandlingen. Om man exempelvis jämför kognitiv beteendeterapi (KBT) med psykodynamisk terapi, så ska man säkerställa att terapeuterna som genomför KBT inte ägnar sig åt moment som hör till psykodynamisk terapi och vice versa. En mer noggrann definition av differentiering är att man för varje intervention i en stu-

die specificerar och mäter (a) unika och nödvändiga komponenter, (b) nödvändiga, men ej unika komponenter, (c) komponenter som varken är unika eller nödvändiga och (d) avrådda/förbjudna komponenter (Dane & Schneider, 1998). Det råder dock viss begreppsförvirring, eftersom vissa använder ovanstående definition när de avser att mäta följsamhet (t.ex. Waltz m.fl., 1993), vilket i så fall gör begreppet differentiering överflödigt.

Steg 3. Finns ett etablerat mätinstrument?

För vissa behandlingar finns det etablerade instrument och procedurer för att mäta behandlingstrohet. Instrument med dokumenterade egenskaper, eller som åtminstone har använts i tidigare studier, är att föredra framför ett eget instrument. Ofta måste emellertid nya instrument utvecklas som passar för den aktuella behandlingen och sammanhanget. När studien av Komet genomfördes fanns inget etablerat instrument att tillgå som rörde behandlarnas följsamhet. Det fanns amerikanska mätinstrument för liknande interventioner, men det specifika innehållet session för session var så pass olikt att ett eget instrument behövde utvecklas. Däremot kunde ett etablerat instrument användas för klientföljsamhet (d.v.s. i vilken grad föräldrarna ändrade sina beteenden). Det instrumentet mätte emellertid generella beteenden, varför föräldrarnas följsamhet vad gäller specifika hemuppgifter i Komet i stället fick ingå i det egenutvecklade instrumentet. I de följande stegen beskrivs hur egna mätinstrument av följsamhet kan utvecklas.

Steg 4a. Utgå från manualen för att formulera kärnkomponenter

Följsamhet handlar om att mäta i vilken grad specificerade delar av en behandling genomförs. Därför är behandlingsmanualen den naturliga utgångspunkten om man ska utveckla ett eget mätinstrument. I princip innebär 100 procent följsamhet att behandlingen implementeras precis enligt manualen. Eftersom det inte går att mäta varje liten detalj i en manual måste innehållet sammanfattas

på en lagom nivå. Övergripande beteckningar som psykodynamisk terapi eller kognitiva strategier är till exempel för allmänna för att kunna mätas tillförlitligt (Garland, Hurlburt, Brookman-Frazee, Taylor & Accurso, 2010). Å andra sidan är det inte heller lämpligt med alltför detaljerade operationaliseringar av innehållet i behandlingen (t.ex. specifika yttranden av behandlaren). Dels kan omfattande mätningar innebära mycket merarbete för behandlare och klienter, dels försämras reliabiliteten i instrument som innehåller alltför många olika dimensioner och moment (Hagermoser Sanetti, Chafouleas, Christ & Gritter, 2009). Garland med flera (2010) refererar till flera forskare som rekommenderar att behandlingens innehåll beskrivs i termer av kärnkomponenter. Kärnkomponenterna formuleras på en nivå som ligger någonstans mellan teoretiska begrepp och specifika yttranden. De kan både syfta på innehåll/aktiviteter där klienten ska agera (t.ex. avslappningsövningar) och strategier som behandlaren använder sig av för att förmedla innehållet eller direkt påverka klienten (t.ex. sokratisk frågeteknik). I tabell 10:2 finns exempel på kärnkomponenter i Komet.

Många artiklar som inkluderar egenutvecklade mått på behandlingstrohet förbiser att beskriva hur man har kommit fram till behandlingens kärnkomponenter och fokuserar i högre grad på hur mätningen skett, till exempel genom beskrivningar av skattnings-

Tabell 10:2. Kärnkomponenter i Komet för föräldrar.

Innehåll/aktivitet	Behandlingsstrategi
• Gemensam lek • Proaktivt beteende (förberedelser) • Positiva uppmaningar • Beröm • Poängsystem hemma • Poängsystem i skolan • Välja strider • Regler • Nödbroms (variant av time-out) • Insyn (t.ex. med vem och var är barnet?) • Problemlösning	• Följa dagordningen • Genomföra rollspel och låta alla föräldrar vara aktiva • Genomföra modellering (filmer) • Ge individualiserade hemuppgifter • Följa upp hemuppgifter och ge individualiserad feedback

skalor och observationer (Mowbray, Holter, Teague & Bybee, 2003). Det är en brist eftersom validiteten i mätningen är helt beroende av att man har gjort en grundlig analys av innehållet i behandlingen. Analysen innebär att enskilda delar av manualen som är särskilt betydelsefulla, omfattande och förväntat effektiva sammanfattas till mer övergripande komponenter. Frågor som kan vägleda analysen är:

- Vilka moment ska behandlarna enligt manualen ägna mycket tid åt? Denna fråga är central. Följsamhet handlar om att mäta vad behandlarna faktiskt gör under behandlingen. Det spelar mindre roll att man tycker att vissa moment är centrala eller viktiga, om de ändå inte är prioriterade tidsmässigt i manualen.

  I Komet ska behandlarna enligt manualen ägna ojämförligt mest tid åt att formulera och följa upp hemuppgifter. Nästan 50 procent av tiden som är avsatt i dagordningarna handlar om hemuppgifter.
- Anger manualen några moment som särskilt viktiga att genomföra?

  Vikten av hemuppgifter betonas starkt både i behandlingsmanualen och i material till föräldrarna. Likaså betonas vikten av att genomföra rollspel (i stället för att berätta/föreläsa).
- Vad förväntas klienterna göra? Finns det specificerade hemuppgifter? Dessa frågor är ett indirekt sätt att komma fram till vilka moment i behandlingen som är centrala. Det som klienterna förväntas göra på egen hand bör rimligen vara viktigt för behandlaren att ägna tid och kraft åt. Det finns starka argument för att även formulera kärnkomponenter (och frågor i mätinstrumentet) som också rör klientföljsamhet. I någon mening speglar det utfall av en behandling: Det spelar ingen roll hur följsamma och kompetenta behandlarna är om det inte resulterar i förändrade beteenden hos klienterna.

  När kärnkomponenterna till Komet formulerades användes just hemuppgifterna till varje session som vägledning. Även om manualen angav att behandlarna skulle genomföra en rad olika moment under en session, så valdes just de moment som sedan resulterade i en hemuppgift ut

till kärnkomponenter (vänsterkolumn i tabell 10:2). Frågor i mätinstrumenten gällde både om behandlarna hade gått igenom respektive moment under sessionerna och om föräldrarna sedan hade genomfört respektive hemuppgift.

- Vilka delar av manualen utgör specifika/unika moment? Kan snarlika moment slås samman? Eftersom alla delar av manualen inte kan mätas måste man göra avgränsningar och prioriteringar. Processen påminner om principerna för faktoranalys. Varje kärnkomponent (eller faktor) ska vara tillräckligt unik för att förtjäna att inkluderas. Samtidigt är strävan att begränsa kärnkomponenter (faktorer) på ett meningsfullt och praktiskt sätt, vilket innebär att närliggande komponenter (faktorer) kan slås samman. När man har kommit så långt i processen att man har ett färdigt mätinstrument kan faktoranalyser användas för att testa hur komponenterna förhåller sig till varandra.

  Kometmanualen innehåller ett antal tydligt avgränsade färdigheter som föräldrarna ska öva på. Kärnkomponenter gällande innehållet i behandlingen var därför relativt lätta att formulera (vänsterkolumn i tabell 10:2). Däremot var det inte lika uppenbart vilka behandlarstrategier som var centrala för att förmedla innehållet (högerkolumn i tabell 10:2). Manualen innehåller många uppmaningar om vad behandlarna ska göra och tänka på. Till sist avgränsades och prioriterades ett fåtal behandlarstrategier som bedömdes vara återkommande genom behandlingen, mätbara och viktiga för att påverka föräldrarnas beteenden.

Steg 4b. Vad gör man med behandlingar som saknar en tydlig manual?

Det finns flera situationer där formulering av kärnkomponenter inte är lika enkel som i exemplet med Komet. Det är till exempel svårare om en manual saknar tydlig struktur och anvisningar om centrala moment. Eller om behandlingen är komplex och omfattar många moment, behandlare, klienter och miljöer. I sådana fall krävs en mer omfattande kvalitativ analys för att bestämma behandlingens kärnkomponenter – vad som ska mätas. Det innebär vanligtvis att

experter anlitas för en bedömning, där en uppenbar svårighet är att avgöra vem som är expert. Mowbray med kollegor (2003) rekommenderar att man använder sig av personer som har olika perspektiv på behandlingen – forskare, praktiker och klienter. Dessutom påtalar de att det finns en tendens hos experter att värdera alla delar i en behandling som viktiga, i synnerhet om experten är involverad i utvecklingen av behandlingen. Det kan därför vara lämpligt att använda sig av tvingande rankingsystem i bedömningen av behandlingens innehåll.

Den här typen av kvalitativa analyser underlättas av att behandlingen bygger på en uttalad teori (Malysiak, Duchnowski, Black & Greeson, 1996). Då ökar möjligheterna att enas om vilka slags moment och behandlingsstrategier som bör prioriteras och som kan antas vara verksamma. Komet utgår till exempel från inlärningsteori och tillämpad beteendeanalys. Det var en hjälp i arbetet att fastslå vilka behandlingsstrategier som skulle prioriteras som kärnkomponenter (högerkolumn tabell 10:2). Ännu bättre är det förstås om det finns empiri som underlag för att bestämma kärnkomponenter. I Komet kan till exempel betoningen av hemuppgifter motiveras av forskning som generellt har påvisat hur viktiga de är för utfall i behandlingar (t.ex. Kazantzis, Whittington & Dattilio, 2010).

Att mäta följsamhet i behandlingar som helt saknar manual innebär en särskild utmaning. Bohart, O'Hara och Leitner (1998) föreslår att man i studier av sådana behandlingar ska försöka hitta bredare formuleringar som rör de generella principer och filosofiska antaganden som behandlingen bygger på snarare än att försöka precisera specifika tekniker. Detaljerade manualer förekommer mer sällan i processorienterade behandlingar (t.ex. psykodynamisk terapi) än i färdighetsbaserade behandlingar (t.ex. KBT). Det är en förklaring till att mätning av följsamhet är vanligare i studier av färdighetsbaserade behandlingar (Perepletchikova & Kazdin, 2005). Svårigheten att mäta följsamhet har emellertid också blivit aktuell inom traditionellt färdighetsbaserade inriktningar som KBT (Perepletchikova, 2009), där det på senare år har skett en utveckling med

behandlingar som vilar mer på övergripande principer än på detaljerade manualer (McHugh, Murray & Barlow, 2009).

Steg 5. Välj typ av mätinstrument

När man har bestämt kärnkomponenterna i en behandling måste man bestämma hur dessa ska mätas. Det kan ske genom observationer, självskattningar eller insamling av befintlig dokumentation (t.ex. dagboksanteckningar om hemuppgifter). Nedan följer en beskrivning av för- och nackdelar med de olika mätmetoderna, som också sammanfattas i tabell 10.3. Idealt används flera oberoende källor för att mäta följsamheten, men det är alltid en avvägning mot den tid och kostnad som mätningen innebär (Schulte m.fl., 2009).

Den metod som brukar betraktas som säkrast för mätning av behandlingstrohet är oberoende observationer (Breitenstein m.fl., 2010). Den främsta nackdelen med metoden är den höga kostnaden och tidsåtgången. Det leder till exempel till att man ofta måste begränsa mätningen av behandlingstrohet till ett slumpmässigt antal tillfällen av behandling. Metoden kräver också att observatörer tränas och jämförs med varandra genom studien för att säkerställa interbedömarreliabilitet. Till sist finns det risk för att observatörens närvaro leder till reaktivitet – det vill säga att behandlaren och klienten ändrar sitt beteende vid de tillfällen som observation pågår. Vissa kan bli mer angelägna om att vara behandlingstrogna, medan effekten för andra kan vara att känna sig kontrollerad och därmed bli mindre behandlingstrogen. Risken för reaktivitet minskar om man använder video, eller ännu hellre ljudinspelning. Andra fördelar med video/ljud är att man säkrare kan observera subtila detaljer i ett samspel och att kontroll av interbedömarreliabilitet underlättas. Samtidigt finns alltid risken att behandlare och klient gör saker som hamnar utanför bild eller utom hörhåll.

Fördelen med självskattningar är framför allt att det är enkelt, tidseffektivt och billigt. Nackdelen är framför allt bristande reliabilitet och validitet, främst överskattningar av den egna följsamheten (DiMatteo, 2004; Noell m.fl., 2005). Av kostnadsskäl valdes själv-

Tabell 10:3. Metoder för att mäta följsamhet i behandlingar.

Metod	Definition	Fördelar	Nackdelar
Självskattning	Skattningar utförda av behandlare och/eller klient.	– Billigt och enkelt. – Kan enkelt anpassas till specifika behandlingar. – Ingen utbildning krävs.	– Problem med reliabilitet och validitet (t.ex. överskattning av egen följsamhet och socialt önskvärda svar).
Observation	Oberoende observatörer utför skattningar av behandlingssessioner. Kan ske "live", med hjälp av video- eller ljudinspelning.	– Objektivt mått. – Möjlighet att bedöma kontextuella faktorer (ej ljudinspelning) – Video- och ljudinspelning medför enklare reliabilitetscheck och möjlighet att observera samma session flera gånger.	– Kostsamt och tidskrävande. – Träningsprocedurer av observatörer måste utformas. – Reaktivitet (lägre med video- och ännu lägre med ljudinspelning). – Video- och ljudinspelning kan uppfattas som integritetskränkande. – Video- och ljudinspelning kan missa relevant information. – Ljudinspelning missar icke-verbal kommunikation.
Befintlig dokumentation	Insamling av dokumentation som ingår i behandlingen.	– Ingen extra arbetsinsats krävs av behandlare och klienter. – Kan bli en rutin för kontinuerlig mätning av behandlingstrohet. – Ofta mått med hög upplevd trovärdighet.	– Den befintliga dokumentationen kan sakna viktiga dimensioner av behandlingstrohet. – Osäker reliabilitet och validitet.

skattning som metod i studien av Komet. Eventuella systematiska fel (bias) kompenserades genom att även föräldrarna fick skatta graden av behandlarnas följsamhet.

Befintlig dokumentation (eng. *permanent products*) kan till exempel handla om protokoll, symtomskattningar eller dagböcker som ingår

som en reguljär del av en behandling. Informationen kvantifieras och kan till exempel användas som en grund för att räkna ut hur stor del av en tänkt behandling som har genomförts. Det är givetvis inte möjligt att mäta alla aspekter av behandlingstrohet genom befintlig dokumentation, men när det går så är det en metod med många fördelar. Den största fördelen är att mätning av behandlingstrohet blir en naturlig del av behandlingen. När en behandlingsmodell eller manual utvecklas bör den befintliga dokumentationen utformas med behandlingstrohet i åtanke (Sheridan, Swanger-Gagné, Welch, Kwon & Garbacz, 2009). Om exempelvis aktivitetsplanering är en kärnkomponent i en depressionsbehandling bör förekomsten av den komponenten ingå i behandlarens och/eller klientens dokumentation. Mätning genom befintlig dokumentation innebär också att risken för reaktiva effekter elimineras, eftersom mätningen inte innebär något som avviker från behandlingen. En annan fördel är att befintlig dokumentation kan användas för att validera annan slags mätning av behandlingstrohet (t.ex. självrapportering). Befintlig dokumentation har framför allt använts inom skolforskning (Schulte m.fl., 2009), men principerna går att generalisera till andra sammanhang. I takt med att datorer används mer, både i traditionella behandlingar (journalsystem etc.) och rena internetbehandlingar, ökar möjligheten att samla in befintlig dokumentation på ett enkelt sätt.

Steg 6. Utveckla kärnkomponenter till specifika frågor

Om man har bestämt sig för att använda självskattning eller observationer som mätmetod för följsamhet återstår att utveckla själva mätinstrumentet. I princip skiljer det sig inte från utveckling av andra slags mätinstrument, som beskrivs i kapitel 6. I det här steget kommer några punkter som är viktiga att tänka på just vid mätning av behandlingstrohet. Som exempel utgår vi från den webbenkät som föräldrarna besvarade efter varje session i Kometstudien. Det är inte ett exempel på perfekt mätning av följsamhet utan används här som utgångspunkt för diskussion. I figur 10:1 visas enkäten som skulle fyllas i efter fjärde sessionen.

Är kärnkomponenterna tillräckligt operationaliserade?

Beteenden och händelser som skattas ska vara möjliga att observera och mäta. Ibland måste man därför operationalisera de formulerade kärnkomponenterna för att kunna formulera lämpliga frågor. I vissa fall kan det räcka med en enda fråga för en kärnkomponent i ett självskattningsformulär eller observationsprotokoll. För kärnkomponenten ”Positiva uppmaningar” i Komet fick föräldrarna till exempel svara på frågan ”Gick behandlarna igenom Positiva uppmaningar under träffen? (ja/nej)”. Ibland måste man emellertid formulera flera frågor för tillförlitlig mätning av en kärnkomponent. I figur 10:1 framgår det att kärnkomponenten ”Ormen” har brutits

1. Ungefär hur många gånger har du genomfört hemuppgiften ”gemensam stund” senaste veckan?
   - ☐ 0
   - ☐ 1
   - ☐ 2
   - ☐ 3
   - ☐ 4
   - ☐ 5
   - ☐ 6
   - ☐ 7 eller fler

2. Ungefär hur ofta har du genomfört hemuppgiften ”förbereda, uppmana och berömma ditt barn” den senaste veckan?
   - ☐ Har inte genomfört hemuppgiften
   - ☐ Enstaka gång under veckan
   - ☐ Några gånger under veckan
   - ☐ En gång varje dag
   - ☐ Två gånger varje dag
   - ☐ Tre eller fler gånger varje dag

3. Fick du bra råd och stöd av gruppledarna när du berättade om hemuppgiften?
   - ☐ Ja, mycket bra
   - ☐ Ja, bra
   - ☐ Ja, ganska bra
   - ☐ Nej, inget eller dåligt stöd
   - ☐ Jag fick inte tillfälle att berätta

4. Förklarade gruppledarna poängsystemet ”Ormen” genom rollspel?
   - ☐ Ja
   - ☐ Nej

5. Gick gruppledarna igenom ”uppdragslista”?
   - ☐ Ja
   - ☐ Nej

6. Gick gruppledarna igenom ”vanliga frågor om Ormen”?
   - ☐ Ja
   - ☐ Nej

7. Ungefär hur många filmer tittade ni på under träffen?
   - ☐ 0
   - ☐ 1
   - ☐ 2
   - ☐ 3 eller fler

8. Deltog du i något rollspel under träffen?
   - ☐ Ja
   - ☐ Nej, men andra i gruppen rollspelade
   - ☐ Vi genomförde inga rollspel under träffen

9. På vilket sätt gick gruppledarna igenom hemuppgifterna till kommande träff?
   - ☐ De anpassade hemuppgiften till varje förälder
   - ☐ De gav samma uppgift till alla föräldrar, men gav tillfälle att ställa frågor och diskutera svårigheter
   - ☐ De gav precis samma uppgift till alla föräldrar utan tillfälle att ställa frågor och diskutera svårigheter
   - ☐ Gruppledarna gav inga hemuppgifter

Figur 10:1. Webbenkät efter fjärde sessionen i Kometstudien.

ner till tre frågor i enkäten (fråga 4–6). Att specifika frågor faktiskt mäter en viss kärnkomponent kan också undersökas genom faktoranalys av skalor, vilket diskuteras vidare i kommande steg.

Skalsteg

En vanlig rekommendation för att minska bias är att använda objektiva kriterier för varje skalsteg som är förankrade i konkreta beteenden (Bond, Becker, Drake & Vogler, 1997). Ibland kan dikotoma variabler (t.ex. förekomst ja/nej) vara den mest tillförlitliga skattningen. Emellertid kräver vissa frågor kontinuerliga skattningar. I sådana fall rekommenderas att man använder tio skalsteg (Hagermoser Sanetti m.fl., 2009). Det var inte aktuellt med så många skalsteg i enkäten i figur 10:1, eftersom svaren är anpassade efter det som förväntas enligt manualen. Exempelvis vore det inte rimligt att fler än 3–4 filmer visas under sessionen.

Undvik bias genom tillåtande frågeformuleringar

Säkerheten i självskattningar ökar om frågorna inleds med ett tillåtande konstaterande (Hagermoser Sanetti m.fl., 2009):

> Det är svårt att exakt följa en manual i varje given situation. Det kan till exempel krocka med andra principer eller åtaganden i ditt arbete. Svara på följande frågor utifrån vilka delar av behandlingen som har varit möjligt att genomföra för din del. Syftet med frågorna är inte att granska ditt arbete utan att få mer kunskap om hur behandlingen fungerar i praktiken. Om du av olika skäl har svårt att göra en bra bedömning av beteendet du ska skatta är det bättre om du avstår från att svara på just den frågan – exempelvis om du inte har haft möjlighet att observera ett visst klientbeteende /. . ./ Frågorna handlar om vad du och din behandlare har gjort de senaste sessionerna. Vad som sker i en behandling kan och bör variera från fall till fall. Det är därför sannolikt att din behandlare inte har genomfört alla de moment som förekommer i frågorna. Det är viktigt att du försöker svara så sanningsenligt som möjligt på alla frågor eftersom den kunskap vi

får från patienterna lägger grunden för att kunna utveckla verksamheten och ge så bra vård som möjligt.

Eftersom de flesta frågorna i figur 10:1 handlar om att föräldrarna ska skatta behandlarnas följsamhet, så fanns det inte så stora skäl för tillåtande formuleringar. Däremot skulle det ha varit bra med sådana formuleringar i frågorna som handlar om hur föräldrarna själva har genomfört hemuppgifterna.

Bygg in självskattningar som en reguljär del av behandlingen

Det finns en risk att kontinuerliga självskattningar i sig påverkar följsamheten, vilket kan jämföras med reaktivitet vid observationer. Det kan till exempel leda till att följsamheten blir bättre i studier som inkluderar självrapportering än när behandlingen implementeras i reguljär verksamhet (Hagermoser Sanetti m.fl., 2009). Det finns en risk att detta uppstod i studien av Komet. Deltagarna i studien randomiserades till gruppbehandling, självhjälpsbehandling eller väntelista. I både grupp- och självhjälpsbetingelsen fick deltagarna svara på enkäter efter varje session av det slag som illustreras i figur 10:1. Det är möjligt att dessa täta uppföljningar bidrog till att föräldrarna arbetade mer aktivt med behandlingen än vad de annars hade gjort – särskilt i självhjälpsgruppen. Ett sätt att hantera den här typen av effekter är att låta självskattningar ingå som en naturlig del i behandlingen (se avsnittet Befintlig dokumentation). Detta utgör en fördel med självskattning framför observationer som mätmetod, eftersom reaktiviteten vid observationer inte lika lätt kan hanteras. I Komets fall har självhjälpsmaterialet vidareutvecklats till en internetbehandling, som håller på att undersökas i en separat studie. Det innebär att den täta uppföljningen numera är inbyggd i behandlingen.

Hellre få frågor ofta, än många frågor sällan

Forskning har visat att skattningar av följsamhet som sker sällan, till exempel månadsvis, är mindre tillförlitliga än mer frekventa skatt-

ningar, till exempel dagligen (Christ, Riley-Tillman & Chafouleas, 2009). Det är därmed generellt bättre att genomföra följsamhetsskattningar med färre frågor som kommer oftare än att ha ett stort frågebatteri som administreras sällan eller i slutet av en behandling. En annan fördel med täta mätningar är att eventuell variation av följsamhet över tid kan spåras. Dessutom ger täta mätningar möjlighet att tidigt upptäcka eventuella brister i följsamheten eller mätprocedurer. I Kometstudien genomfördes föräldraskattningarna av följsamhet efter varje session, vilket medförde att specifika frågor kunde ställas om den aktuella sessionens innehåll. Därmed hölls antalet frågor per tillfälle på en begränsad nivå.

Ibland är det bättre med övergripande frågor

Formuleringen av beteenden ska inte vara alltför specifik. Exempelvis kan det vara bättre att ställa en övergripande fråga, som ”Hur stor del av klientens hemuppgifter följdes upp på ett noggrant sätt?”, än att ställa specifika frågor om enskilda hemuppgifter (Riley-Tillman, Chafouleas, Christ, Briesch & LeBel, 2009). I Kometstudien användes en övergripande fråga om just hemuppgifter, eftersom utformningen av hemuppgifter anpassas individuellt för föräldrarna. Därmed var det svårt att ställa specifika frågor, som kanske inte skulle omfatta alla tänkbara beteenden som är relevanta. Samtidigt får frågorna givetvis inte vara alltför övergripande så att beteenden som i hög grad är oberoende av varandra blandas. Exempelvis bör ”Följde upp hemuppgift” och ”Gav ny hemuppgift” skattas separat, snarare än att de klumpas ihop till ”Arbetade med hemuppgifter”.

Skattningen ska ske i nära anslutning (tid och plats) till beteendet som observeras

Vad som är i nära anslutning måste avgöras från fall till fall. Det kan till exempel vara efter varje session, efter varje skoldag, efter varje lektion eller dagliga klientskattningar. Om ett behandlingsmoment omfattar flera miljöer och tidpunkter, bör en skattning ske för varje miljö/tidpunkt. I komplexa behandlingar som omfattar många

olika moment kan de beteenden som skattas variera över tid (d.v.s. matchas mot de aktuella behandlingskomponenterna).

Välj observationsmetod

Om observationer ska användas finns det flera olika tillvägagångssätt att välja mellan. Kontinuerliga skattningar är ofta mer tillförlitliga än skattningar i efterhand. Kontinuerlig skattning innebär att observatören löpande noterar när olika beteenden förekommer (t.ex. hur många gånger en lärare uppmuntrar en elev). Kontinuerliga skattningar kan också innebära så kallad time sampling, som betyder att förekomsten av olika beteenden skattas med täta tidsintervall (t.ex. var 30:e sekund). Den metoden är lämplig för beteenden som förekommer med hög frekvens (t.ex. hur stor del av tiden som en elev har suttit på sin plats). Vissa aspekter av behandlingstrohet är svåra att mäta med kontinuerliga skattningar. När observatören till exempel ska bedöma i vilken grad en behandlare har visat engagemang, så måste bedömningen ske i efterhand (t.ex. på en skala 1–10). Likaså kan skattning i efterhand vara motiverat då en observatör ska rapportera hur stor del av en planerad dagordning som har genomförts under en session. I avsnittet om kompetens längre fram i kapitlet finns ett mer utförligt exempel på hur observationsprotokoll kan utvecklas och tillämpas.

Steg 7. Etablera reliabilitet och validitet för instrumentet

Oavsett om man utvecklar ett eget instrument för att mäta behandlingstrohet eller om man använder ett etablerat instrument, så måste man säkerställa reliabilitet och validitet. Dessvärre används ofta instrument med bristfällig dokumentation i dessa avseenden (Baer m.fl., 2007). För att påminna om inledningen av detta kapitel, så är det anmärkningsvärt att så lite uppmärksamhet ägnas åt att säkerställa tillförlitliga mått av den oberoende variabeln (behandlingen) i effektstudier. Den bristen gäller även Kometstudien, där det i förväg inte fanns någon kontroll av reliabilitet eller validitet i mätinstrumenten. I en kommande artikel kommer analyser av interbedömar-

reliabilitet mellan föräldrar och behandlare samt andra fördjupande analyser av behandlingstrohet i Kometstudien att publiceras.

Med tanke på hur sällan behandlingstrohet kontrolleras så är det en styrka att över huvud taget kunna bifoga något slags information om det. Nedan sker en kort summering av några typer av reliabilitet och validitet som är särskilt centrala i detta sammanhang.

Validitet

Först och främst är det nödvändigt att etablera god innehållsvaliditet (eng. *content validity*). Det handlar om att formuleringar av kärnkomponenter och frågor/item rent begreppsligt ska motsvara det faktiska innehållet i behandlingen. Innehållsvaliditeten kan inte direkt undersökas kvantitativt, utan bygger på den kvalitativa process som beskrevs i det fjärde steget i detta avsnitt. Det går helt enkelt inte att räkna fram vilka komponenter som utgör essensen i en behandling. Det kan bara ske genom en bedömning av något slags experter. Sedan går det förstås att göra kvantitativa beräkningar som rör samstämmigheten i de tillfrågade experternas bedömningar (t.ex. Lawshe, 1975). God innehållsvaliditet erhålls om flera experter med olika perspektiv kan enas om formuleringarna. Innehållsvaliditeten blir också starkare om formuleringar revideras upprepade gånger, med möjlighet till feedback däremellan.

Vid sidan av innehållsvaliditet är det bra att visa att mätinstrumentet kan skilja mellan grupper av behandlare/klienter som borde skilja sig åt. Ett sätt att göra det är att se om behandlare med olika grad av utbildning eller handledning skiljer sig åt i behandlingstrohet enligt det instrument som testas. En annan variant kan vara att jämföra behandlingstrohet för olika typer av klienter, där man till exempel kan förvänta sig lägre följsamhet med mer komplexa fall eller omotiverade klienter.

Prediktiv validitet (eng. *predictive validity*) är också en typ av validitet som brukar diskuteras i samband med behandlingstrohet. Eftersom själva poängen med god behandlingstrohet är att klienterna ska vinna på det är det en klar styrka att kunna visa på sam-

band mellan instrumentet och ett gott utfall i behandlingen. Om man däremot inte finner signifikanta samband, så betyder det inte automatiskt att det är mätinstrumentet för behandlingstrohet som är bristfälligt. Avsaknaden av samband kan också bero på att behandlingen i sig inte är tillräckligt effektiv eller på att det finns brister i hur man mätt utfallet. Därför är det i regel bättre att prioritera undersökningar av instrumentets förmåga att faktiskt skilja mellan bättre och sämre behandlare. När man vet att instrumentet kan göra det, så kan man gå vidare och titta på prediktiv validitet med större säkerhet. Svårigheten att tolka prediktiv validitet berörs ytterligare i ett senare avsnitt om samband mellan behandlingstrohet och utfall.

I studien av Komet förekom ett exempel på prediktiv validitet utan att först ha etablerat att instrumentet kunde diskriminera mellan behandlare som skilde sig åt i följsamhet. Effekterna i studien var större för familjer som deltog i gruppbehandlingen än för dem med självhjälp. Den skillnaden kunde statistiskt helt förklaras av att föräldrarna i gruppbetingelsen genomförde fler hemuppgifter (Kling m.fl., 2010). Därmed uppvisade mätinstrumentet som användes (figur 10:1) god prediktiv validitet.

Reliabilitet

Eftersom det är vanligt att behandlingstrohet mäts med hjälp av observationer, så är interbedömarreliabilitet (eng. *inter-rater reliability*) av särskilt intresse i detta sammanhang. För att undersöka interbedömarreliabilitet skattar minst två oberoende observatörer samma sekvens. Interbedömarreliabilitet kan givetvis också undersökas för självskattningar, där skattningar från olika källor kan jämföras (t.ex. mellan klient och behandlare).

En vanlig undersökning av mätinstrument brukar vara att titta på den inre konsistensen (eng. *internal consistency*). Det innebär att man kan kontrollera huruvida frågor som avser att mäta samma dimension (t.ex. en viss kärnkomponent) faktiskt hänger ihop. Ett problem med inre konsistens är att det är lättare att få bra värden om många frågor används. Ett annat problem är att det inte går att

beräkna inre konsistens om det bara finns en eller ett par frågor som rör en viss dimension. Ett alternativ till att mäta inre konsistens är att genomföra faktoranalys, som kan vara lämplig om instrumentet omfattar många frågor som mäter separata kärnkomponenter (faktorer). Faktoranalysen kan sägas vara en kvantitativ kontroll av operationaliseringen av kärnkomponenterna. Analysen ger vägledning om vilka frågor som hör ihop (d.v.s. hör till en viss kärnkomponent) och vilka frågor som är överflödiga.

Mätning av kompetens

För behandlingar med komponenter som inte går att genomföra på så många olika sätt kan det vara tillräckligt att mäta följsamhet. Vid studier av mer komplexa behandlingar bör man emellertid försöka mäta både följsamhet och kompetens. Medan följsamhet brukar referera till den omfattning med vilken olika delar av en behandling har genomförts, så brukar kompetens definieras som den kvalitet med vilken behandlaren genomför behandlingen. Av flera skäl är det svårare att mäta kompetens än följsamhet. För det första är det svårare att bestämma kriterier för kompetens. För följsamhet kan man ofta utgå från manualer eller annan dokumentation för att bestämma kriterier, medan kriterier för kompetens ofta måste grundas på konsensusbeslut av experter och praktiker. För det andra är det lättare att entydigt avgöra om en behandlare faktiskt har använt en viss teknik (följsamhet) än att avgöra i vilken grad tekniken genomfördes på ett kompetent sätt. För det tredje kräver mätning av kompetens i princip observationsskattningar (som genomförs av ”experter”), eftersom självskattningar av kompetens inte är tillförlitliga (Miller, Yahne, Moyers, Martinez & Pirratano, 2004). Eftersom det inte är givet vem som räknas som expert för en viss behandling bör urvalet av dessa beskrivas så transparent som möjligt (Barber m.fl., 2007).

Följsamhet brukar beskrivas som en grundläggande förutsättning för kompetens. Det går att genomföra en behandling med hög grad

av följsamhet på ett icke-kompetent sätt, men däremot är det omvända inte möjligt (Perepletchikova & Kazdin, 2005). Ett kompetent genomförande av en behandling förutsätter en hög grad av följsamhet, även om den inte är total – eftersom ett kompetent genomförande av en behandling ofta innebär att behandlaren gör vissa avsteg från manualen. Kompetenta avsteg innebär att behandlaren tar hänsyn till kontextuella faktorer (t.ex. klientens situation) och att avstegen sker i överensstämmelse med övergripande teorier och principer i behandlingen. Varje avvikelse från den planerade behandlingen ska kunna motiveras med att de gagnar klienten, till skillnad från icke-kompetenta avsteg som till exempel sker av bekvämlighetsskäl eller på grund av bristande färdigheter hos behandlaren.

Vid mätning av kompetens kan både globala och specifika aspekter inkluderas. Globala aspekter kan till exempel vara skattningar av generell kvalitet på genomförandet av en session och/eller i vilken grad sessionens innehåll anpassades efter klientens behov (Houge m.fl., 2008). För specifika aspekter av kompetens kopplas skattningarna till enskilda delar av behandlingen. Ett exempel på det senare är *The Therapist Behavior Rating Scale - Competence* (TBRS-C), där observatörer skattar följsamhet och kompetens för varje kärnkomponent i behandlingen (Houge m.fl., 2008). Jämfört med behandlingar som sker under kontrollerade förhållanden (t.ex. terapirum), så kan det vara svårare att uppnå reliabilitet i kompetensskattningar i behandlingar som genomförs i mer okontrollerade sammanhang (t.ex. skola eller hem). Det gäller särskilt vid skattningar av specifika dimensioner, vilket gör att globala skattningar kan vara att föredra ju lägre grad av kontroll som föreligger (Houge m.fl., 2008).

Barber med flera (2007) uttrycker tvivel över möjligheten att tillmötesgå de ökande kraven på att mäta och skilja ut kompetens från följsamhet. De beskriver till exempel hur olika instrument brister vad gäller psykometriska egenskaper, distinktion mellan kompetens och följsamhet samt precisering av global/specifik kompetens. Trots dessa brister påtalar de vikten av fortsatt strävan mot att definiera

och mäta kompetens, som annars bara blir en del av felkällan i behandlingstroheten.

Ett norskt exempel på mätning av kompetens

I två norska studier av föräldraträning för barn med utagerande problem (PMTO) har man genomfört mätningar och analyser av behandlarnas kompetens (Forgatch & DeGarmo, 2011; Ogden & Amlund-Hagen, 2008). Mätinstrumentet som användes kallas för *Fidelity of Implementation Rating System* (*FIMP)* (Knutson, Forgatch & Rains, 2003). Instrumentet bygger på bedömningar av videoinspelade behandlingssekvenser. I studierna användes två 10-minuterssekvenser från två olika sessioner som grund för bedömningen. Utbildade PMTO-terapeuter med goda kunskaper om principer, kärnkomponenter och behandlingstekniker genomförde bedömningarna. De fick utbildning och träning i FIMP grundad på den manual som hör till instrumentet. Utgångspunkten för att kunna genomföra bedömningar är att interbedömarreliabiliteten överstiger 70 procent. Bedömningarna omfattar fem dimensioner som ska spegla i vilken grad behandlarna utgår från innehållet i manualen samtidigt som de tar hänsyn till familjernas behov, situation och reaktioner. Var och en av dimensionerna bedöms på en niogradig skala, där 1–3 innebär icke-godkänt, 4–6 godkänt och 7–9 väl godkänt. De fem dimensionerna är:

1. KUNSKAP OM PMTO. Behandlaren tillämpar behandlingsmodellens principer, använder rätt tekniker och procedurer och förstår grundläggande föräldrafärdigheter.
2. STRUKTUR. Behandlaren följer en plan för sessionen, driver behandlingen framåt – har flyt, leder utan att styra, är lyhörd för ämnen som familjen tar upp, tar hänsyn till timing och sammanfattar vid lämpliga tillfällen.
3. UNDERVISNING. Den verbala undervisningen ska omfatta pedagogiska metoder som att informera, komma med förslag och att förklara grunden för behandlingen (ge rational). Den aktiva undervisningen ska omfatta brainstorming, rollspel och att ställa frågor

till familjerna. Behandlaren ska visa att han eller hon behärskar strategier som medför att familjerna på ett självständigt sätt kan tillämpa innehållet i behandlingen.

4. PROCESS. Behandlaren ska sträva efter att etablera ett tryggt och stödjande sammanhang under sessionerna för att främja behandlingsprocessen. Färdigheterna omfattar till exempel att ställa öppna frågor, möta familjers motstånd och att hantera konflikter. Det handlar om att behärska flera terapeutiska verktyg som exempelvis användning av metaforer, humor, reflektioner, ”spegling” och ”reframing”.
5. KVALITET. Denna dimension är mer övergripande och överlappar delvis övriga dimensioner. Här bedöms den generella kvaliteten i behandlingen, vilken baseras på en rad olika faktorer. Exempelvis omfattar dimensionen hur väl behandlaren genomför specifika inslag i behandlingen, hur väl mål med sessionerna uppnås samt i vilken grad föräldrarna verkar tillämpa och uppskatta behandlingen.

Etablerade mätinstrument av behandlingstrohet

Det här avsnittet beskriver ett urval av mätinstrument som har använts i behandlingsstudier. Mätinstrument med dokumenterade psykometriska egenskaper är naturligtvis att föredra framför att utveckla ett helt eget instrument. Även om det inte finns ett instrument som exakt passar den aktuella studien eller behandlingen, så kanske något eller några av instrumenten i avsnittet kan vara en grund för utvecklingen av egna instrument. I genomgången nedan framgår det att instrumenten har dokumenterade psykometriska egenskaper, men när man tittar närmare på källorna så är det tydligt att i princip varje instrument också har brister. Exempelvis kan påståenden om ”god interbedömarreliabilitet” dölja avsevärd variation i reliabilitet mellan delskalor i instrumentet. Det samlade intrycket av litteraturen är att det fortfarande krävs ytterligare forskning för att med säkerhet kunna rekommendera ett visst mätinstrument.

Kognitiv terapi

Kognitiv terapi är en behandlingsform där det finns relativt många studier som har inkluderat mätning av behandlingstrohet. En av de mest använda skalorna är *Cognitive Therapy Scale* (CTS; Young & Beck, 1980), som bygger på observatörsskattningar av videoinspelade sessioner. Den mäter dimensionen kompetens, även om det kan ifrågasättas om den inte samtidigt inkluderar element som rör följsamhet (Barber m.fl., 2007). Observatörer utgår från en manual som omfattar 11 aspekter av kompetens som omfattar både generella terapeutfärdigheter (t.ex. *Collaboration*) och färdigheter specifika för kognitiv terapi (t.ex. *Guided discovery*). Instrumentet har god inre konsistens och interbedömarreliabiliteten har visat sig vara tillfredsställande (Vallis, Shaw & Dobson, 1986). Skalan diskriminerar mellan sessioner som enligt oberoende skattare uppfyller respektive inte uppfyller kraven på god kognitiv terapi (Vallis m.fl., 1986). CTS-poäng har också visat sig predicera symtomminskning i behandlingsstudier (Kingdon, Tyrer, Seivewright, Ferguson & Murphy, 1996; Trepka, Rees, Shapiro, Hardy & Barkham, 2004), även om det inte gäller för alla utfallsmått (Shaw m.fl., 1999).

Ett instrument som inspirerats av CTS är *Cognitive Therapy Adherence and Competence Scale* (CTACS; Liese, Barber & Beck, 1995). Instrumentet bygger på oberoende skattningar av ljudinspelningar av sessioner och omfattar 25 item (t.ex. ”*Elicited automatic thoughts and related these to patient's problems*”). För varje item skattas både följsamhet och kompetens, vilket skiljer denna skala från många andra som använder olika item eller skalor för följsamhet respektive kompetens. I en undersökning av instrumentet påvisades psykometriska egenskaper som i vissa avseenden överträffade CTS (Barber, Liese & Abrams, 2003).

Psykodynamisk och interpersonell psykoterapi

En skala som är specifikt utformad för psykodynamisk korttidsterapi är *Penn Adherence/Competence Scale for Supportive-Expressive Dynamic Psychotherapy* (PACS-SE; Barber, 1988). Den påminner om CTACS

i strukturen, eftersom både följsamhet (specificerat som frekvens av aktiviteter) och kompetens (specificerat som kvalitet och lämplighet i tillämpningen) skattas för varje item. Totalt omfattar skalan 45 item som är indelade i tre teoretiskt härledda delskalor; *general therapeutic skills*, *supportive skills* och *expressive skills*. Den sistnämnda innehåller item som är mest specifika för psykodynamisk terapi, exempelvis ”The therapist focuses attention on similarities among the patient’s past and present relationships”. Förutom god inre konsistens och interbedömarreliabilitet har skalan visat sig diskriminera mellan psykodynamisk och kognitiv terapi (Barber & Crits-Christoph, 1996). I en studie av relationen mellan behandlingstrohet och behandlingsutfall med deprimerade patienter fann forskarna att terapeuternas följsamhet inte hade någon betydelse (Barber, Crits-Christoph & Luborsky, 1996). Däremot fanns det ett tydligt samband mellan en delskala för kompetens (*expressive skills*) och utfall, även när man tog hänsyn till alternativa förklaringar som generell terapeutkompetens och allians.

Motiverande samtal (MI)

Motiverande samtal (eng. *motivational interviewing*) är en behandlingsmodell som framför allt använts med klienter som har problem med alkohol eller droger. Det finns flera instrument för att bedöma behandlingstrohet som har använts i studier av motiverande samtal, varav *Motivational Interviewing Treatment Integrity Code* (MITI) är ett av de mest använda (Moyers, Martin, Manuel, Hendrickson & Miller, 2005). Instrumentet bygger på ljudinspelningar av sessioner och omfattar flera olika dimensioner. Skattare bedömer följsamhet genom att räkna antalet gånger som behandlaren utför vissa beteenden. Även differentiering bedöms i viss mån, eftersom skattaren noterar när ”förbjudna” beteenden förekommer. Till sist bedöms kompetens genom att fem variabler (t.ex. *Collaboration* och *Empathy*) bedöms på en femgradig skala. Instrumentet har god interbedömarreliabilitet och har visat känslighet för att mäta effekter av träning av behandlare (Moyers m.fl., 2005; Mitcheson, Bhavsar

& McCambridge, 2009). Motsvarande psykometriska egenskaper har dokumenterats i svenska studier av MITI (Forsberg, Kallmen, Hermansson, Berman & Helgason, 2007; Forsberg, Berman, Kallmen, Hermansson & Helgason, 2008).

Parent Management Training (PMT)

FIMP (Fidelity of Implementation Rating System) är ett instrument som mäter behandlarnas kompetens vid genomförande av Parent Management Training (Knutson m.fl., 2003). Innehållet i FIMP beskrevs i avsnittet om kompetens. Den prediktiva validiteten har undersökts i en mindre amerikansk (Forgatch, Patterson & DeGarmo, 2005) och en omfattande norsk studie (Ogden, Forgatch, Askeland, Patterson & Bullock, 2005; Forgatch & DeGarmo, 2011). I båda studierna fanns linjära samband mellan FIMP och de mått som mätte föräldrarnas sätt att uppmuntra och sätta gränser för barnen. I den norska studien prövades också om det fanns ett samband mellan FIMP och det slutgiltiga utfallsmåttet – barnens beteenden (Forgatch & DeGarmo, 2011). Trots att forskargruppen i en rad tidigare studier har funnit samband mellan föräldrars och barns beteenden, så fann man i denna studie inte något direkt samband mellan FIMP och barnens beteenden.

Multisystemisk terapi (MST)

Multisystemisk terapi är en behandlingsmodell som riktar sig till ungdomar med utagerande problematik (Henggeler, Schoenwald, Borduin, Rowland & Cunningham, 2009). Det finns en inbyggd mätning av följsamhet i modellen som kallas för *Therapist Adherence Measure* (TAM; Henggeler, Borduin, Schoenwald, Huey & Chapman, 2006). Instrumentet har använts i många studier men i olika versioner, där antalet item har varierat mellan 15 och 28. Det bygger på att klienterna blir uppringda en gång i månaden och får skatta i vilken grad deras terapeuter har genomfört olika delar av behandlingen och följt nio formulerade behandlingsprinciper. En finess med instrumentet är att även terapeuterna skattar sina hand-

ledare med ett motsvarande instrument. Man har funnit att klienternas skattningar av terapeuternas följsamhet prediceras av terapeuternas skattningar av sina handledare, vilket stödjer validiteten i instrumentet (Schoenwald, Sheidow & Letourneau, 2004; Schoenwald, 2008).

Även diskriminativ validitet har demonstrerats i en studie där man fann signifikanta skillnader i TAM-skattningar när MST jämfördes med en annan behandling (Henggeler m.fl., 2006). En av de främsta styrkorna med TAM är den prediktiva validiteten, som har påvisats i flera studier genom samband mellan följsamhet och behandlingsutfall (t.ex. återfall i brott). Exempelvis visar Schoenwald, Chapman, Sheidow och Carter (2009) i en långtidsuppföljning av nästan 2000 klienter att terapeuternas följsamhet förklarar en stor del av återfallsfrekvens i brott, när man samtidigt kontrollerar för en lång rad andra variabler. Liknande samband har även påvisats i andra amerikanska studier (Huey, Henggeler, Brondino & Pickrel, 2000; Schoenwald, Carter, Chapman & Sheidow, 2008) samt en norsk studie (Ogden & Halliday-Boykins, 2004). Däremot saknades samband mellan TAM och utfall i en svensk studie av MST, vilket fick författarna att ifrågasätta validiteten i måttet i svenska sammanhang (Sundell m.fl., 2008).

Instrument som omfattar flera slags behandlingar

Det finns ett antal exempel på studier där man använt instrument som mäter behandlingstrohet i flera slags behandlingar samtidigt. Ett av de mest kända är *Yale Adherence and Competence Scale* (YACS; Carroll m.fl., 2000). Det utvecklades för att mäta trohet i tre olika behandlingar för missbruksproblem: 12-stegsprogrammet, KBT och kliniskt omhändertagande (*clinical management*). Det omfattar 55 item, där observatörer skattar både följsamhet och kompetens (kvalitet) i videoupptagningar av sessioner. Skalan bygger på differentiering, det vill säga den omfattar både moment som ingår respektive inte ingår i de olika behandlingarna. Carroll med kollegor (2000) fann god interbedömarreliabilitet och god inre konsistens.

Till skillnad från en del andra skalor så var det svaga till måttliga samband mellan dimensionerna för följsamhet och kompetens, vilket styrker att de faktiskt mäter olika saker. Man fann även stöd för diskriminativ validitet, eftersom det fanns signifikanta skillnader mellan relevanta delskalor för de tre behandlingarna.

En relativt välanvänd skala är *Collaborative Study Psychotherapy Rating Scale* (CSPRS; Hollon m.fl., 1988). Skalan mäter följsamheten och omfattar dimensioner som rör KBT, interpersonell psykoterapi (ITP) och kliniskt omhändertagande (*clinical management*). Den utvecklades för behandlingsstudier med deprimerade klienter, men har även använts för till exempel anorexi och missbruksproblem (Baranackie, Crits-Christoph & Kurcias, 1992; McIntosh m.fl., 2005). Skalan bygger på observatörsskattningar av ljudinspelade sessioner och den omfattar totalt 96 item. God interbedömarreliabilitet och inre konsistens har dokumenterats (DeRubeis & Feeley, 1990; McIntosh m.fl., 2005). Skalan diskriminerar väl mellan behandlingarna (Beranackie m.fl., 1992; McIntosh m.fl., 2005). Även prediktiv validitet har demonstrerats genom samband mellan delskalor i CSPRS och behandlingsutfall (DeRubeis & Feeley, 1990).

Ett instrument som skiljer sig något från övriga i denna genomgång är *Comparative Psychotherapy Process Scale* (CPPS; Hilsenroth, Blagys, Ackerman, Bonge & Blais, 2005). Det är inte utvecklat för behandlingar av ett visst problem eller en viss population, utan mäter följsamhet i två olika behandlingar, oberoende av klientgrupp. Skalan omfattar 20 item och kan användas för självskattning av terapeuter/klienter eller av oberoende observatörer. Den omfattar två delskalor, psykodynamisk-interpersonell och kognitiv-beteendeterapeutisk. Innehållet i delskalorna bygger på litteraturgenomgångar av terapeutbeteenden som är karakteristiska för behandlare som arbetar psykodynamiskt, interpersonellt, kognitivt och beteendeterapeutiskt – såväl unika och essentiella som ”förbjudna” item ingår i skalan. I en studie av Hilsenroth med flera (2005), där skalan presenteras i sin helhet, påvisades mycket goda psykometriska egenskaper. De fann hög interbedömarreliabilitet, god inre konsi-

stens och starka samband med liknande skalor. De visade också att instrumentet hade god diskriminativ validitet genom stora och signifikanta skillnader mellan olika behandlingar i de skattningar som genomfördes av observatörer, handledare, terapeuter och klienter.

Samband mellan behandlingstrohet och utfall

En livligt debatterad fråga de senaste decennierna är om behandlingstrohet har någon betydelse för utfallet av behandlingen. Vissa hävdar att det är generella faktorer (t.ex. allians) snarare än det specifika behandlingsinnehållet som spelar roll (t.ex. Wampold, 2001). Andra menar att det specifika innehållet visst har betydelse, men att det inte finns några entydiga mönster vad gäller sambandet mellan trohet och utfall (t.ex. Barber m.fl., 2007; McHugh m.fl., 2009; Perepletchikova & Kazdin, 2005). Vissa studier har till exempel funnit positiva linjära samband; ju bättre behandlingstrohet, desto bättre effekt (t.ex. Frank & Frank, 1991; Huey m.fl., 2000; Kling m.fl., 2010; McHugo, Drake, Teague & Xie, 1999; Schoenwald m.fl., 2009). Andra studier har funnit kurvlinjära samband för följsamhet som tyder på att en optimal tillämpning av en behandling inte är samma sak som att följa manualen till punkt och pricka (t.ex. Hogue m.fl., 2008; Barber m.fl., 2006). Flera studier har inte funnit några samband alls (t.ex. Barlow, Gorman, Shear & Woods, 2000; Loeb m.fl., 2005; Mihalic, 2004; Sundell m.fl., 2008) och några har till och med funnit negativa samband (t.ex. Castonguay, Goldfried, Wiser, Raue & Hayes, 1996; Huppert m.fl., 2006). Negativa samband eller icke-samband behöver dock inte betyda att behandlingen är ineffektiv.

Ska behandlingstrohet betraktas som ett generellt fenomen?

Vad är vanligast? Positiva, negativa eller inga samband alls? I en aktuell metaanalys av 36 studier som undersökt effekterna av behandlingstrohet fann man inte några sammanvägda signifikanta samband mellan behandlingstrohet och utfall (Webb, DeRubeis & Barber,

2010). Sambandet mellan följsamhet och behandlingsutfall var r = .02 och motsvarande samband för kompetens var r = .07. Den här typen av sammanvägda resultat tas ibland som intäkt just för att det specifika innehållet i behandlingar saknar betydelse (Wampold, 2001). Förutom att sådana tolkningar ofta bortser från metodologiska brister i de ingående studierna (Bhar & Beck, 2009), så bygger de också på en föreställning om att behandlingstrohet är ett universellt fenomen som har relevans oberoende av behandlingsform, population och studiedesign. Webb med flera (2010) tar i diskussionen av resultaten upp att det var stor variation mellan de ingående studierna och att en uniform effekt vore oväntad. Om man i en behandlingsstudie kan visa att effekten är signifikant större för de deltagare som får mer/bättre intervention, så tyder det på att det är det specifika innehållet i interventionen som ligger bakom effekten (Kazdin, 2007). Ett sådant resultat försvagas inte av att behandlingstrohet inte har visat några effekter för andra slags behandlingar. Problem uppstår först om effekter av behandlingstrohet på samma slags behandling visar motstridiga reslutat i olika studier. Det är därför inte lämpligt att betrakta behandlingstrohet som ett generellt begrepp. Det bör i stället betraktas som ett begrepp som i viss mån är unikt för varje enskild studie, eftersom specifika behandlingsmodeller, populationer och metoder för att mäta behandlingstrohet varierar så mycket. Exempelvis har metaanalyser som begränsats till en viss typ av behandling och sammanhang (beteendemodifikation i skolan) funnit signifikanta samband mellan trohet och utfall (Noell m.fl., 2005; Gresham, Gansle, Noell, Cohen & Rosenblum, 1993). Ett annat exempel är en metaanalys av betydelsen av hemuppgifter (d.v.s. klientföljsamhet) i KBT-behandling (Kazantzis m.fl., 2010). När författarna kontrasterade samma slags behandling med och utan hemuppgifter visade analysen en tilläggseffekt på Cohens d = .48. Ytterligare ett exempel är en metaanalys av program för elever med utagerande problematik (Wilson & Lipsey, 2007); problem med implementering (d.v.s. bristande behandlingstrohet) var en av de faktorer som hade starkast betydelse för effektstorleken i analysen.

Metodologiska problem och lösningar i analyser av behandlingstrohet

Vid sidan av problemet att klumpa ihop resultat från olika studier konstaterar Perepletchikova och Kazdin (2005) att det finns en lång rad metodologiska frågor som man måste ta hänsyn till. En grundläggande förutsättning för att kunna finna signifikanta samband mellan behandlingstrohet och utfall är att reliabla och valida mått har använts. Detta problem är särskilt stort gällande behandlarens kompetens, där man till exempel ofta har haft svårt att uppnå interbedömarreliabilitet mellan skattare (Barber m.fl., 2007). Avsaknad av samband mellan behandlingstrohet och utfall kan också bero på oklara definitioner och bristfällig validitet – det vill säga att de skalor och observationer som används helt enkelt inte mäter de relevanta aspekterna av en behandling (Perepletchikova & Kazdin, 2005). Specifika problem som rör validitet kan vara oklar definition av behandlingstrohet (innehållsvaliditet) och en alltför avgränsad definition av behandlingstrohet som enbart tar hänsyn till enstaka dimensioner av begreppet (Webb m.fl., 2010). Ett specifikt problem som rör reliabilitet kan vara att behandlingstroheten enbart mäts vid enstaka sessioner, vilket maskerar en vanligt förekommande variation av trohet över tid i en behandling (Webb m.fl., 2010). Ett återkommande problem vid analyser av behandlingstrohet är takeffekter. Eftersom man ofta försöker optimera behandlingstrohet i interventionsstudier, genom noggrann träning och uppföljning av behandlare, så tenderar behandlingstroheten att vara generellt hög. Det kan också förklara varför det har varit lättare att hitta samband i studier av behandling som sker i reguljära verksamheter, eftersom variationen i behandlingstrohet mellan behandlare generellt blir större då (McHugh m.fl., 2009).

I de fall då man enbart undersöker linjära samband mellan behandlingstrohet och utfall (t.ex. Webb m.fl., 2010), så finns risken att man missar relevanta icke-linjära samband. Ett exempel på det här problemet är relationen mellan följsamhet och kompetens. Följsamhet och kompetens är förstås beroende av varandra; en kompetent genomförd behandling förutsätter till exempel en viss grad av

följsamhet. Samtidigt finns en motsättning mellan begreppen. Hög terapeutkompetens kan till exempel innebära att behandlingen anpassas efter enskilda individer, vilket kan leda till avsteg från en föreskriven behandling och därmed lägre grad av följsamhet (McHugh m.fl., 2009). För att mått på behandlingstrohet ska vara valida måste den typen av dynamiska effekter beaktas. Till exempel kan man behöva precisera vilka avsteg från en behandlingsmanual som är möjliga utan att inverka negativt på behandlingstroheten som helhet.

I många analyser av behandlingstrohet beaktas inte effekten av modererande faktorer, som exempelvis klientegenskaper. Om sådana faktorer förbises kan effekter av behandlingstrohet maskeras. Till exempel visade Ellis, Weiss, Han och Gallop (2010) att en rad klientfaktorer påverkade behandlarnas följsamhet i tillämpningen av multisystemisk terapi. Följsamheten var bättre med familjer som hade starkare sammanhållning, högre förväntningar på behandlingen och som var inställda på att behandlingen skulle innebära att de själva måste göra förändringar. Följsamheten påverkades också av föräldrarnas psykiska hälsa och deras inställning till uppfostran. Det här är ett exempel på att behandlare anpassar tillämpningen av en insats beroende på klientens egenskaper och svårigheter. Behandlingstroheten kan både öka och minska som följd av svårare klientfall (Webb m.fl., 2010). Det är förstås inte bara klientfaktorer som kan moderera effekterna av behandlingstrohet. Andra faktorer, som till exempel rör behandlarens organisation, kan också påverka. Ett exempel är Payne (2009) som visade att en rad organisatoriska variabler i skolor påverkade behandlingstroheten vid genomförande av både individuella och skolomfattande insatser för elever. Därmed inte sagt att organisationsfaktorer har en generell betydelse för alla slags behandlingar. Schoenwald med flera (2009) visade till exempel att organisationsfaktorer i princip inte alls hade någon effekt på återfall i brottslighet hos klienterna när de analyserades tillsammans med behandlarens följsamhet.

Till sist, för att kunna dra säkra slutsatser om effekter av behandlingstrohet bör man samtidigt ta hänsyn till andra variabler som rör

behandlingsprocessen. Det kan ske genom att man kontrollerar för effekten av sådana variabler i analyserna, eller genom att de inkorporeras i definitioner av behandlingstrohet. Exempelvis bör man kontrollera för betydelsen av allians mellan behandlaren och klienten, alternativt betrakta allians som en del av behandlingstroheten (Sharpless & Barber, 2009).

Sammanfattning

Här sammanfattas de viktigaste punkterna att tänka på när behandlingstrohet ska preciseras, mätas och analyseras i en studie.

1. **Modellutvärdering eller verksamhetsbaserad utvärdering.** Om det är en modellutvärdering som genomförs, så är det önskvärt att mäta och optimera behandlingstroheten innan förmätningen genomförs. Om det är en studie av behandlingens effekter i reguljär verksamhet mäts den faktiska behandlingstroheten utan att extra medel eller resurser tillförs.
2. **Ska studien jämföra flera aktiva behandlingar?** I så fall kan det vara aktuellt att mäta differentiering. Det innebär att man säkerställer att inslag som hör till behandling A inte används i behandling B och vice versa.
3. **Prioritera följsamhet.** Följsamhet är en grundläggande dimension av behandlingstrohet som bör prioriteras – åtminstone om det handlar om manualbaserade behandlingar.
4. **Använd etablerade skalor.** Använd etablerade mätinstrument om det finns något att tillgå som passar behandlingen i fråga.
5. **Utveckla egna mätinstrument.** Om ett eget mätinstrument ska utvecklas bör följande steg beaktas:
   a. *Formulera kärnkomponenter*. Vilka är de essentiella inslagen i behandlingen; vad är viktigt att mäta? Utgångspunkten i arbetet är manualer och annan dokumentation som beskriver innehållet i behandlingen. Experter och helst behandlare och klienter konsulteras för att identifiera essentiella delar. Med fördel utgår processen från den teoretiska och empiriska grunden för

behandlingen. I behandlingar som saknar detaljerade manualer är processen mer komplicerad och expertutlåtanden av olika slag får större betydelse.

b. *Välj form av mätning*. Mätningar av behandlingstrohet kan ske genom olika slags självskattningar, observationer eller genom insamling av dokumentation som förekommer naturligt i behandlingen. Det finns för- och nackdelar med alla mätmetoder. Idealfallet är därför att genomföra flera parallella mätningar.

c. *För självskattningar och observationer: formulera frågor*. För att mätningen ska bli meningsfull måste en bra operationalisering av kärnkomponenterna genomföras. Generellt är det bra att använda relativt få frågor som man ställer ofta i självskattningar. Det blir ofta mer tillförlitligt än instrument som innehåller många frågor och som distribueras sällan eller vid enstaka tillfällen. Likaså är kontinuerliga observationer av beteenden ofta mer tillförlitliga än observatörsskattningar som sker i efterhand. Skattningarna vinner på om de sker kort tid efter de aktuella behandlingsmomenten.

d. *Etablera validitet*. Undersök i första hand instrumentets förmåga att skilja mellan bättre och sämre behandlare. Det är en styrka att dessutom kunna visa prediktiv validitet, men kan samtidigt vara svårt att tolka avsaknad av signifikanta samband.

e. *Etablera reliabilitet*. Kontrollera inre konsistens och genomför eventuellt faktoranalys för att kontrollera operationaliseringarna av kärnkomponenterna. Om observationer ska genomföras krävs kontroll av interbedömarreliabilitet, något som också kan vara aktuellt om självskattningar genomförs med flera källor (t.ex. information från behandlare och klienter).

6. **Överväg mätning av kompetens.** Kompetens är en dimension av behandlingstrohet som är särskilt relevant att mäta i mer komplexa behandlingar, som just ställer stora krav på att behandlarna genomför behandlingen på ett skickligt sätt. Det är emellertid ofta svårare att mäta kompetens än att mäta följsamhet. Det är också ofta mer tidskrävande, eftersom det i regel krävs observationer

med expertbedömningar. Flera instrument för mätning av behandlingstrohet som har utvecklats på senare år låter observatörer skatta både följsamhet och kompetens samtidigt. Det innebär att man först bedömer i vilken grad behandlaren har genomfört ett behandlingsmoment och sedan skattar hur skickligt det i så fall genomfördes.

7. **Samband mellan behandlingstrohet och utfall.** Det är komplicerat att analysera samband mellan behandlingstrohet och utfall i en behandling. Det finns många felkällor att beakta och det krävs hög grad av kontroll av ovidkommande och modererande variabler för att kunna dra säkra slutsatser. Det är heller inte säkert att sambanden är linjära, varför mer komplexa analyser ibland är nödvändiga. Kort sagt kan man säga att påvisande av samband innebär en styrka både för mätinstrumentet av behandlingstrohet (prediktiv validitet) och för slutsatserna om att behandlingen i fråga är effektiv (dos-respons-samband). Problemen uppstår framför allt när det inte finns några tydliga samband, eftersom det är svårt att ha kontroll över alla felkällor.

Fördjupningslitteratur

Barber, J. P., Sharpless, B. A., Klosterman, S. & McCarthy, K. S. (2007). Assessing intervention competence and its relation to therapy outcome: a selected review derived from the outcome literature. *Professional Psychology: Research and Practice, 38*, 493–500.

Kazdin, A. E. (2007). Mediators and moderators of change in psychotherapy research. *Annual Review of Clinical Psychology, 3,* 1–27.

Mowbray, C. T., Holter, M. C., Teague, G. B. & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation, 24,* 315–340.

Perepletchikova, F. & Kazdin, A. E. (2005). Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice, 12,* 365–383.

Schulte, A. C., Easton, J. E. & Parker, J. (2009). Advances in treatment integrity research: Multidiciplinary perspectives on the conceptualization, measurement, and enhancement of treatment integrity. *School Psychology Review, 38,* 460–475.

Referenser

Baer, J. S., Ball, S. A., Campbell, B. K., Miele, G. M., Schoener, E. P. & Tracy, K. (2007). Training and fidelity monitoring of behavioral interventions in multisite addictions research. *Drug and Alcohol Dependence, 87*, 107–118.

Baranackie, K., Crits-Christoph, P. & Kurcias, J. S. (1992). Therapist techniques used during the cognitive therapy of opiate-dependent patients. *Journal of Substance Abuse Treatment, 9,* 221–228.

Barber, J. P. (1988). *The Penn Adherence/Competence Scale for Supportive-Expressive Psychotherapy*. Opublicerat manus. Center for Psychotherapy Research, Department of Psychiatry, University of Pennsylvania.

Barber, J. P. & Crits-Christoph, P. (1996). Development of an adherence/ competence scale for dynamic therapy: Preliminary findings. *Psychotherapy Research, 6,* 81–94.

Barber, J. P., Crits-Christoph, P. & Luborsky, L. (1996). Effects of therapist adherence and competence on patient outcome in brief dynamic therapy. *Journal of Consulting and Clinical Psychology, 64,* 619–622.

Barber, J. P., Liese, B. & Abrams, M. J. (2003). Development of the Cognitive Therapy Adherence and Competence Scale. *Psychotherapy Research, 13,* 205–221.

Barber, J. P., Gallop, R., Crits-Christoph, P., Frank, A., Thase, M., Weiss, R. D., m.fl. (2006). The role of therapist adherence, therapist competence, and alliance in predicting outcome of individual drug counseling: results from the National Institute on Drug Abuse Collaborative Cocaine Treatment Study. *Psychotherapy Research, 16*, 229–240.

Barber, J. P., Sharpless, B. A., Klosterman, S. & McCarthy, K. S. (2007). Assessing intervention competence and its relation to therapy outcome: a selected review derived from the outcome literature. *Professional Psychology: Research and Practice, 38*, 493–500.

Barlow, D. H., Gorman, J. M., Shear, M. K. & Woods, S. W. (2000). Cognitive behavioral therapy, imipramine, or their combination for panic disorder: a randomized controlled trial. *Journal of the American Medical Association, 283*, 2529–2536.

Bhar, S. S. & Beck, A. T. (2009). Treatment integrity of studies that compare short-term psychodynamic psychotherapy with cognitive-behavior therapy. *Clinical Psychology: Science and Practice, 16*, 370–378.

Bohart, A. C., O'Hara, M. & Leitner, L. M. (1998). Empirically violated treatments: Disenfranchisement of humanistic and other psychotherapies. *Psychotherapy Research, 8*, 141–157.

Bond, G. R., Becker, D. R., Drake, R. E. & Vogler, K. M. (1997). A fidelity scale for the individual placement and support model of supported employment. *Rehabilitation Counseling Bulletin*, *40*, 265–284.

Breitenstein, S. M., Gross, D., Garvey, C. A., Hill, C., Fogg, L. & Resnick, B. (2010). Implementation fidelity in community-based interventions. *Research in Nursing and Health, 33,* 164-173.

Carroll, K. M., Nich, C., Sifry, R. L., Nuro, K. F., Frankforter, T. L., Ball, S. A., m.fl. (2000). A general system for evaluating therapist adherence and competence in psychotherapy research in the addictions. *Drug and Alcohol Dependence, 57,* 225–238.

Castonguay, L. G., Goldfried, M. R., Wiser, S., Raue, P. J. & Hayes, A. M. (1996). Predicting the effect of cognitive therapy for depression: A study of unique and common factors. *Journal of Consulting and Clinical Psychology, 64,* 497–504.

Christ, T. J., Riley-Tillman, T. C. & Chafouleas, S. M. (2009). Foundation for the development and use of direct behavior rating (DBR) to assess and evaluate child behavior. *Assessment for Effective Intervention, 34,* 201–213.

Dane, A. V. & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review, 18,* 23–45.

DeRubeis, R. J. & Feeley, M. (1990). Determinants of change in cognitive therapy for depression. *Cognitive Therapy and Research, 14*, 469–482.

DiMatteo, M. R. (2004). Variations in patients' adherence to medical recommendations: A quantitative review of 50 years of research. *Medical Care*, *42*, 200–209.

Ellis, M. L., Weiss, B., Han, S. & Gallop, R. (2010). The Influence of Parental Factors on Therapist Adherence in Multi-systemic Therapy. *Journal of Abnormal Child Psychology, 38*, 857–868.

Forgatch, M. S. & DeGarmo, D. (2011). Sustaining fidelity following the nationwide PMTO implementation in Norway. *Prevention Science, 12,* 235–246.

Forgatch, M. S., Patterson, G. R. & DeGarmo, D. S. (2005). Evaluating fidelity: Predictive validity for a measure of competent adherence to the Oregon model of parent management training. *Behavior Therapy, 36*, 3–13.

Forsberg, L., Kallmen, H., Hermansson, U., Berman, A. H. & Helgason, A. R. (2007). Coding counsellor behaviour in motivational interviewing sessions: inter-rater reliability for the Swedish Motivational Interviewing Treatment Integrity Code (MITI). *Cognitive Behaviour Therapy, 36*, 162–169.

Forsberg, L., Berman, A. H., Kallmén, H., Hermansson, U. & Helgason, A. R. (2008). A test of the validity of the motivational interviewing treatment integrity code. *Cognitive Behavior Therapy, 37,* 183–191.

Frank, J. D. & Frank, J. B. (1991). *Persuasion and healing: A comparative study of psychotherapy* (3rd ed.). Baltimore, MD: Johns Hopkins University Press.

Garland, A. F., Hurlburt, M. S., Brookman-Frazee, L., Taylor, R. M. & Accurso, E. C. (2010). Methodological Challenges of Characterizing Usual Care Psychotherapeutic Practice. *Adm Policy Ment Health, 37*, 208–220.

Gresham, F. M., Gansle, K., Noell, G. H., Cohen, S. & Rosenblum, S. (1993). Treatment integrity of school-based behavioral intervention studies: 1980–1990. *School Psychology Review, 22*, 254–272.

Hagermoser Sanetti, L. M., Chafouleas, S. M., Christ, T. J. & Gritter, K. L.

(2009). Extending DBR use beyond student assessment: Applications to treatment integrity assessment within a multi-tier model of school-based intervention delivery. *Assessment for Effective Intervention, 34*, 251–258.
Henggeler, S. W., Borduin, C. M., Schoenwald, S. K., Huey, S. J. & Chapman, J. E. (2006). *Multisystemic Therapy Adherence Scale – Revised (TAM-R).* Charleston, SC: Medical University of South Carolina, Department of Psychiatry and Behavioral Science.
Henggeler, S. W., Halliday-Boykins, C. A., Cunningham, P. B., Randall, J., Shapiro, S. B. & Chapman, J. E. (2006). Juvenile drug court: Enhancing outcomes by integrating evidencebased treatments. *Journal of Consulting and Clinical Psychology, 74*, 42–54.
Henggeler, S. W., Schoenwald, S. K., Borduin, C. M., Rowland, M. D. & Cunningham, P. B. (2009). *Multisystemic therapy for antisocial behavior in children and adolescents.* New York: Guilford.
Hilsenroth, M. J., Blagys, M. D., Ackerman, S. J., Bonge, D. R. & Blais, M. A. (2005). Measuring psychodynamic-interpersonal and cognitive-behavioral techniques: Development of the Comparative Psychotherapy Process Scale. *Psychotherapy: Theory, Research, Practice, Training, 42,* 340–356.
Hogue, A., Henderson, C. E., Dauber, S., Barajas, P. C., Fried, A. & Liddle, H. A. (2008). Treatment adherences, competence, and outcome in individual and family therapy for adolescent behavior problems. *Journal of Consulting and Clinical Psychology, 76*, 544–555.
Hollon, S. D., Evans, M. D., Auerbach, A., DeRubeis, R. J., Elkin, I., Lowery, A., Kriss, M., Grove, W., Thason, V. B. & Piasecki, J. (1988). *Development of a system for rating therapies for depression: Differentiating cognitive therapy, interpersonal therapy, and clinical management pharmacotherapy.* (Opublicerat manus, University of Minnesota,Twin Cities Campus.)
Huey, S. J., Henggeler, S. W., Brondino, M. J. & Pickrel, S. G. (2000). Mechanisms of change in multisystemic therapy: Reducing delinquent behavior through therapist adherence and improved family and peer functioning. *Journal of Consulting and Clinical Psychology, 68*, 451–467.
Huppert, J. D., Bufka, L. F., Barlow, D. H., Gorman, J. M., Shear, M. K. & Woods, S. W. (2006). The interaction of motivation and therapist adherence predicts outcome in cognitive-behavioral therapy for panic disorder: preliminary findings. *Cognitive and Behavioral Practice, 13*, 198–204.
Kaunitz, C. & Strandberg, A. (2009). Aggression Replacement Training (ART) i Sverige – evidensbaserad socialtjänst i praktiken? *Socionomen, 9,* (26), 37–50.
Kazantzis, N., Whittington, C. & Dattilio, F. (2010). Meta-analysis of homework effects in cognitive and behavioral therapy: A replication and extention. *Clinical Psychology Science and Practice, 17,* 144–156.
Kazdin, A. E. (2007). Mediators and moderators of change in psychotherapy research. *Annual Review of Clinical Psychology, 3,* 1–27.
Kingdon, D., Tyrer, P., Seivewright, N., Ferguson, B. & Murphy, S. (1996). The

Nottingham Study of Neurotic Disorder: Influence of cognitive therapists on outcome. *British Journal of Psychiatry, 169,* 93–97.
Kling, Å., Forster, M., Sundell, K. & Melin, L. (2010). A Randomized Controlled Effectiveness Trial of Parent Management Training with varying degree of therapist support. *Behavior Therapy, 41,* 530–542.
Knutson, N. M., Forgatch, M. S. & Rains, L. A. (2003). *Fidelity of Implementation Rating System (FIMP): The training manual for PMTO.* Eugene, Oregon Social Learning Center.
Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563–575.
Liese, B. S., Barber, J. P. & Beck, A. T. (1995). *The Cognitive Therapy Adherence and Competence Scale.* Opublicerat manus, University of Kansas Medical Center.
Loeb, K. L., Wilson, G. T., Labouvie, E., Pratt, E. M., Hayaki, J., Walsh, B. T., m.fl. (2005). Therapeutic alliance and treatment adherence in two interventions for bulimia nervosa: a study of process and outcome. *Journal of Consulting and Clinical Psychology, 73*, 1097–1107.
Malysiak, R., Duchnowski, A., Black, M. & Greeson, M. (1996). Establishing wrap around fidelity through participatory evaluation. In *Proceedings of the Ninth Annual Research Conference. A System of Care for Children's Mental Health: Expanding the Research Base*.
McHugh, R. K., Murray, H. W. & Barlow, D. H. (2009). Balancing fidelity and adaptation in the dissemination of empirically-supported treatments: The promise of transdiagnostic interventions. *Behaviour Research and Therapy, 47*, 946–953.
McHugo, G. J., Drake, R. E., Teague, G. B. & Xie, H. (1999). Fidelity to assertive community treatment and client outcomes in the new Hampshire dual disorders study. *Psychiatric Services, 50*, 818–824.
McIntosh, V. V. W., Jordan, J., McKenzie, J. M., Luty, S. E., Carter, F. A., Carter, J. D., Frampton, C. M. A. & Joyce, P. R. (2005). Measuring therapist adherence in psychotherapy for anorexia nervosa: Scale adaptation, psychometric properties, and distinguishing psychotherapies. *Psychotherapy Research, 15*, 339–344.
Mihalic, S. (2004). The importance of implementation fidelity. *Emotional and Behavioral Disorders in Youth, 4,* 83–105.
Miller, W. R., Yahne, C.E., Moyers, T. B., Martinez, J. & Pirratano, M. (2004). A randomized trial of methods to help clinicians learn motivational interviewing. *Journal of Consulting and Clinical Psychology, 72*, 1050–1062.
Mitcheson, L., Bhavsar, K. & McCambridge, J. (2009). Randomized trial of training and supervision in motivational interviewing with adolescent drug treatment practitioners. *Journal of Substance Abuse Treatment, 37,* 73–78.
Mowbray, C. T., Holter, M. C., Teague, G. B. & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation, 24,* 315–340.

Moyers, T. B., Martin, T., Manuel, J. K., Hendrickson, S. M. L. & Miller, W. R. (2005). Assessing competence in the use of motivational interviewing. *Journal of Substance Abuse Treatment, 28,* 19–26.

Naleppa, M.J. & Cagle, J.G. (2010). Treatment Fidelity in Social Work Intervention Research: A Review of Published Studies. *Research on Social Work Practice*. (Först publicerat 27 januari 2010 on-line: doi:10.1177/1049731509352088.)

Noell, G. H., Witt, J. C., Slider, N. J., Connell, J. E., Gatti, S. L., Williams, K. L., m.fl. (2005). Treatment implementation following behavioral consultation in schools: A comparison of three follow-up strategies. *School Psychology Review, 34*, 87–106.

Ogden, T. & Amlund-Hagen, K. (2008). Treatment effectiveness of Parent Management Training in Norway: A randomized controlled trial of children with conduct problems. *Journal of Consulting and Clinical Psychology, 76,* 607–621.

Ogden, T., Forgatch, M. S., Askeland, E., Patterson, G. R. & Bullock, B. M. (2005). Implementation of parent management training at the national level: The case of Norway. *Journal of Social Work Practice, 19,* 317–329.

Ogden, T. & Halliday-Boykins, C. (2004). Multisystemic treatment of antisocial adolescents in Norway: Replication of clinical outcomes outside of the US. *Child and Adolescent Mental Health, 9*, 77–83.

Payne, A. A. (2009). Do Predictors of the Implementation Quality of School-Based Prevention Programs Differ by Program Type? *Prevention Science, 10*, 151–167

Perepletchikova, F. & Kazdin, A. E. (2005). Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice, 12,* 365–383.

Perepletchikova, F., Treat, T. & Kazdin, A. E. (2007). Treatment integrity in psychotherapy research: Analysis of the studies and examination of the associated factors. *Journal of Consulting and Clinical Psychology, 75*, 829–841.

Perepletchikova, F. (2009). Treatment integrity and differential treatment effects. *Clinical Psychology: Science and Practice, 16,* 379–382.

Riley-Tillman, T. C., Chafouleas, S. M., Christ, T. J., Briesch, A. M. & LeBel, T. J. (2009). The impact of wording and behavioral specificity on the accuracy of direct behavior ratings (DBRs). *School Psychology Quarterly, 24,* 1–12.

Schoenwald, S. K., Sheidow, A. S. & Letourneau, E. J. (2004). Toward effective quality assurance in evidence-based practice: Links between expert consultation, therapist fidelity, and child outcomes. *Journal of Child and Adolescent Clinical Psychology, 33,* 94–104.

Schoenwald, S. K. (2008). Toward evidence-based implementation of evidence-based treatments: MST as an example. *Journal of Child and Adolescent Substance Abuse, 17*, 69–91.

Schoenwald, S. K., Carter, R. E., Chapman, J. E. & Sheidow, A. J. (2008). Therapist adherence and organizational effects on change in youth behavior problems one year after Multisystemic Therapy. *Administration and Policy in Men-*

tal Health and Mental Health Services Research, 35, 379–394.
Schoenwald, S. K., Chapman, J. E., Sheidow, A. J. & Carter, R. E. (2009). Long-Term Youth Criminal Outcomes in MST Transport: The Impact of Therapist Adherence and Organizational Climate and Structure. *Journal of Clinical Child & Adolescent Psychology, 38*, 91–105.
Schulte, A. C., Easton, J. E. & Parker, J. (2009). Advances in treatment integrity research: Multidiciplinary perspectives on the conceptualization, measurement, and enhancement of treatment integrity. *School Psychology Review, 38,* 460–475.
Sharpless, B. A. & Barber, J. P. (2009). A conceptual and empirical review of the meaning, measurement, development, and teaching of intervention competence in clinical psychology. *Clinical Psychology Review, 29*, 47–56.
Shaw, B. F., Elkin, I., Yamaguchi, J., Olmstead, M., Vallis, T. M., Dobson, K. S., m.fl. (1999). Therapist competence ratings in relation to clinical outcome in cognitive therapy of depression. *Journal of Consulting and Clinical Psychology, 67,* 837–846.
Sheridan, S. M., Swanger-Gagné, M., Welch, G. W., Kwon, K. & Garbacz, S. A. (2009). Fidelity Measurement in Consultation: Psychometric Issues and Preliminary Examination. *School Psychology Review, 38,* 476–495.
Sundell, K., Hansson, K., Andrée Löfholm, C., Olsson, T., Gustle, L-H. & Kadesjö, C. (2008). The transportability of a multisystemic therapy to Sweden: short term results from a randomized trial of conduct-disordered youths. *Journal of Family Psychology, 22,* 550–560.
Trepka, C., Rees, A., Shapiro, D. A., Hardy, G. E. & Barkham, M. (2004). Therapist competence and outcome of cognitive therapy for depression. *Cognitive Therapy and Research, 28*, 143–157.
Vallis, T. M., Shaw, B. E. & Dobson, K. S. (1986). The Cognitive Therapy Scale: Psychometric properties. *Journal of Consulting and Clinical Psychology, 54,* 381–385.
Waltz, J., Addis, M. E., Koerner, K. & Jacobson, N. S. (1993). Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. *Journal of Consulting and Clinical Psychology, 61*, 620–630.
Wampold, B. E. (2001). *The Great Psychotherapy Debate: Models, Methods and Findings.* Mahwah, NJ: Lawrence Erlbaum.
Webb, C. A., DeRubeis, R. J. & Barber, J. P. (2010). Therapist Adherence/Competence and Treatment Outcome: A Meta-Analytic Review. *Journal of Consulting and Clinical Psychology, 78*, 200–211.
Wilson, S. J. & Lipsey, M. W. (2007). School-based interventions for aggressive and disruptive behavior. Update of a meta-analysis. *American Journal of Preventive Medicine, 33,* 130–143.
Young, J. E. & Beck, A. T. (1980). *The development of the Cognitive Therapy Scale*. Opublicerat manus, University of Pennsylvania, Philadelphia, PA.

Kapitel 11 | Bassam Michel El-Khouri

Att hantera bortfall[1]

Data i forskningsstudier som involverar kvantitativ metodik är så gott som aldrig helt kompletta. Förlorad information kan föra med sig allvarliga konsekvenser för de analyser som är planerade. Om bortfallet ignoreras kan det medföra att studiens grundläggande förutsättningar inte längre håller, vilket påverkar tillförlitligheten i slutsatserna. Bortfall kan också leda till att hela analysenheter med enbart partiella data utesluts med följden att viktig information, som antagligen hade bidragit till mer tillförlitliga slutsatser, går förlorad. Bortfall som inte kan eller bör ignoreras kan hanteras under diverse antaganden som när de inte är uppfyllda i sig kan leda till ny bias. Med andra ord: man vinner något på bekostnad av något annat. Detta kapitel tar upp aspekter på analyser där data analyseras enligt den ursprungliga planen för studien; analyser enligt ITT (”intention-to-treat”). ITT-analyser är särskilt relevanta i randomiserade kontrollerade studier (RCT). Kapitlet innehåller även en kortfattad inblick i metoder som används för att hantera bortfall genom imputering.

1 Översättningen från engelska till svenska har gjorts av Åsa Kling, institutionen för psykologi, Uppsala universitet, samt författaren själv.

Intention-To-Treat – ITT

ITT-analyser är en metodologisk konsekvens av att välja en randomiserad kontrollerad design för en studie, den design som många anser vara den främsta för att jämföra eller testa effekter av en intervention. Som namnet antyder innebär en randomiserad kontrollerad studie i sin enklaste form att individer randomiseras, det vill säga fördelas slumpmässigt, till en behandlingsgrupp som får den avsedda behandlingen, medan övriga får en alternativ behandling (t.ex. placebo) eller ingen behandling alls (t.ex. väntelista). Randomisering används för att garantera att alla faktorer utöver utfallsmåtten fördelar sig i genomsnitt lika i grupperna. Oavsett om dessa faktorer mäts eller inte, eller för den delen är mätbara eller inte, säkerställer randomisering att varje meningsfull skillnad i utfallet är en effekt av behandlingen ifråga och inte av någon eller några av de andra faktorerna som kunde tänkas ha en effekt på utfallet. Oavsett hur deltagarna förhåller sig till dessa andra faktorer har de samma sannolikhet att få behandling, att få alternativ behandling eller att inte få någon behandling alls. Om randomisering inte har använts finns det inget som säkerställer att den observerade behandlingseffekten inte är resultatet av att deltagare i en viss grupp har mer eller mindre av någon annan faktor än själva behandlingen och ingen möjlighet att säkert skilja ut den effekten från en eventuell behandlingseffekt. I korthet är randomisering nödvändig för att skapa grupper med lika fördelning av olika karakteristika som kan påverka utfallet för att därigenom minimera systematiska fel i uppskattningen av behandlingseffekter.

Även om randomisering är det främsta instrumentet för att säkerställa en tillförlitlig uppskattning (eng. *unbiased estimate*) av behandlingseffekter krävs det mer. Till det hör att förvissa sig om att utfallsmåtten har samlats in på ett korrekt och tillförlitligt sätt för alla deltagare och att bortfall hanterats på ett sådant sätt att det inte snedvrider jämförelserna mellan behandlings- och kontrollgrupp. För att minimera systematiska fel så långt det är möjligt används,

när det är möjligt, en dubbelblind strategi där vare sig deltagare eller behandlare får veta vilken behandling en deltagare får eller vilka som värderar utfallet. Sättet att hantera bortfall av data kan däremot vara mer eller mindre komplext och kan ha allvarlig inverkan på vilket resultat som erhålls i jämförelser mellan grupper.

I forskningsstudier används ofta den ovannämnda strategin för att undvika olika typer av systematiska fel, medan strategierna för analys kan vara ganska olika. Formellt sett är en ITT-strategi nödvändig för att behålla effekterna av en randomisering för studiens slutsatser. Den möjliggör en ändamålsenlig hantering av deltagare som avbrutit medverkan i förtid. Det betyder att oavsett om deltagare var följsamma eller inte, fick den behandling de skulle eller inte, om de av misstag fick fel behandling, om de mådde dåligt av behandlingen och drog sig ur, om de helt avbröt medverkan i studien eller inte medverkade i all datainsamling, eller av vilken annan anledning som helst avvek från den ursprungliga planen, ska alla bedömningar av utfallet göras som det var planerat och alla deltagare ska inkluderas i jämförelser mellan grupperna samt analyseras enligt den grupptillhörighet de ursprungligen randomiserats till.

Syftet med ITT

ITT syftar till att spåra effekter av interventioner i praktiken, inte effekterna för dem som varit följsamma och avslutat den avsedda behandlingen. Det är en viktig distinktion, för om data från icke följsamma (deltagare som inte följer behandlingsplanen) utelämnas, eventuellt på grund av praktiska skäl som uppges efter randomiseringen, kan analyserna inte längre sägas mäta den praktiska effektiviteten med behandlingen. ITT är avsett för studier av ”den pragmatiska effektiviteten” med en behandling, där man garderar sig mot att överskatta behandlingseffektiviteten. Motsatsen är analyser som endast inkluderar följsamma deltagare (de som följt den ursprungliga behandlingsplanen), en strategi som kallas TOT (”treatment on the treated”). Studien är då explanatorisk, den undersöker den faktiska behandlingseffektiviteten för en subgrupp deltagare och inte

interventionens inverkan när den tillämpas i praktiken. Analyser som utesluter alla icke följsamma deltagare kallas ibland även för per protokoll-analyser (eng. *per protocol*), enligt behandlingen (eng. *on treatment, by treatment administered*) eller explorativa analyser. Resultaten från sådana analyser anses allmänt ha större risk att vara snedvridna: systematiska fel i dessa ökar risken för antalet falska positiva utfall, särskilt i stora studier.

Användning av ITT-analyser är inte alltid utan komplikationer. Effektstorleken på en effektiv intervention eller behandling beräknad endast på följsamma deltagare kan exempelvis underskattas om ITT-analys används och icke-följsamheten är omfattande. Om samtliga deltagare ingår i jämförelserna försvagas analysernas statistiska styrka (eng. *power*) och behandlingseffekten kan bli försvagad, i synnerhet om många deltagare är icke-följsamma (eng. *noncompliant*). Därför är det viktigt att minimera icke-följsamhet i en studie, exempelvis genom att försöka identifiera icke-följsamma deltagare före randomiseringen och försäkra sig om att data är fullständiga med avseende på de primära utfallsmåtten genom att följa icke-följsamma deltagare till studiens slut.

Ofullständig uppföljning underminerar principerna för ITT och reducerar det till en selektiv urvalsanalys. Om primära utfallsdata fortfarande saknas ska imputering användas för att uppskatta eller predicera de saknade värdena, baserat på antaganden om data och typen av bortfall. Det är viktigt att notera att imputerade data är kvalificerade gissningar baserade på antaganden som oftast inte går att verifiera, vilket gör att de slutsatser som dras utifrån dessa data är svagare än data från kompletta dataset.

Avhopp

Det finns en mängd olika anledningar till avhopp (eng. *dropping out*). Deltagare kan ha flyttat eller emigrerat och går inte att nå, de kan ha inkluderats felaktigt, självsvåldigt ha bytt behandlingsgrupp eller inte vill svara eller ställa upp vid uppföljningsmätningar. Ofta

skiljer man mellan avhopp från en behandling och avhopp från en studie. I ITT-ansatsen finns redan inbyggt att avhopp från behandlingen inte är tillräckligt för att exkludera deltagare från studien.

Det finns dock situationer där kraven på att inkludera alla deltagare kan lättas något. Ett klassiskt exempel när det kan vara befogat att utesluta eller exkludera deltagare är då deltagare hoppar av studien efter randomisering men innan behandlingen har startat. I det fallet är det nödvändigt att tiden mellan randomisering och behandlingens start är densamma i alla behandlingsgrupper.

Även i studier som utvärderar deltagarnas följsamhet före randomisering kan det vara berättigat att utesluta icke-följsamma deltagare. Under alla omständigheter måste studiens slutsatser väga in de förutsättningar och avsikter som gäller för studien ifråga.

En annan situation handlar om falska inkluderingar, att individer inkluderas som efter randomiseringen visar sig inte uppfylla inklusionskriterierna för studien. Generellt är det inte ett tillräckligt skäl för att utesluta dem. Mot detta har argument framförts att om fastställelse av inklusionskriterier har tillämpats identiskt i varje undersökningsalternativ kan uteslutning vara befogad. Detta kan dock ifrågasättas eftersom falska inkluderingar ofta förekommer i praktiken och om uteslutning då tillämpas försvagar det möjligheten att generalisera resultat från sådana studier.

Exempel där det är befogat med uteslutning kan handla om individers säkerhet, om det exempelvis finns anledning att befara att igångsättning av eller en fortsatt behandling kan riskera en persons fysiska eller psykiska hälsa. I sådana situationer är det tillrådligt att använda en exkluderingsprocedur som är blindad för deltagarnas behandlingstillhörighet för att gardera sig mot systematiska exklusionsfel mellan behandlingsgrupper.

Jämfört med TOT resulterar ITT i en konservativ uppskattning av behandlingseffekten på grund av att icke-följsamma deltagare ingår i analyserna. ITT behåller randomiseringen och gör det möjligt att generalisera resultat till den avsedda populationen i motsats till TOT-analys, som är inriktad på behandlingseffekter bland

dem som fullföljt behandlingen. Detta för att ITT ger icke snedvridna uppskattningar (eng. *unbiased estimates*) av behandlingseffekter i motsats till uppskattningar av effekter baserade på TOT, där risken är större för systematiska fel. I synnerhet gäller detta om deltagare som hoppat av svarar sämre på behandlingen och om det finns en koppling mellan att följa behandlingen och vilken behandlingsgrupp man ingår i.

Genom att använda en ITT-strategi kontrolleras för faktorer som påverkar följsamhet i grupperna så länge som randomiseringen bibehålls och grupperna blir jämförbara vid baslinjen. Om randomiseringen inte bibehålls riskerar grupperna att inte vara jämförbara vid förmätningen och effekten av faktorer som påverkar fullföljande av behandlingen förblir okänd. En explorativ analys eller TOT-analys kan komplettera ITT-analysen för att bättre förstå kliniska, terapeutiska eller prognostiska skillnader som annars inte skulle komma fram.

Om deltagare utesluts efter randomisering ökar risken för typ I-fel; nollhypotesen förkastas när den i verkligheten är sann. Det innebär att en uppmätt skillnad mellan en behandlings- och kontrollgrupp godtas när ingen sådan skillnad finns i verkligheten. Om inte det uteslutna bortfallet är helt slumpmässigt, blir de återstående data snedvridna till fördel för en skillnad mellan grupperna, eftersom avhopp från ursprungsplanen ofta hänger ihop med behandlingen.

Rekommendationer

Det finns många rekommendationer i litteraturen om principer för analyser av ITT. Om en ITT-ansats är valet har det fundamental betydelse för bland annat studiens design, genomförande, analys av data och rapportering av resultatet. En fördjupad diskussion av dessa områden i relation till ITT ligger utanför ramarna för detta kapitel, men allmänna rekommendationer har getts av Hollis och Campbell (1999):

ITT bör betraktas som en komplett strategi för design, genomförande och analys av en studie snarare än som enbart en strategi för analys.

Design

- Bestäm om målet med studien är pragmatiskt (eng. management trials) eller explanatoriskt (eng. explanatory trial). I pragmatiska studier är ITT grundläggande.
- Motivera i förväg varje inklusionskriterie som när den inte uppfylls tillåter att utesluta deltagare i en ITT-analys.

Genomförande

- Minimera bortfall från primära utfallsmått.
- Följ upp deltagare som hoppar av behandlingen.

Analys

- Inkludera alla deltagare i de grupper de randomiserats till.
- Undersök möjlig effekt av bortfall.

Resultatrapportering

- Ange att ITT-analys har använts samt beskriv hur avvikelser från randomisering och bortfall har hanterats.
- Rapportera avvikelser från randomisering och bortfall.
- Diskutera möjlig effekt av bortfallet.
- Basera slutsatser på resultaten från ITT-analysen.

Imputering

Mycket arbete har lagts ner på att definiera bortfall av data och på att etablera ändamålsenliga metoder för att hantera det. I det här avsnittet diskuteras bortfall inom ramen för randomiserade kontrollerade studier och ITT. Flera av de strategier och imputeringsmetoder som presenteras här är dock tillämpbara även med andra designer.

Inledningsvis konstaterades att ITT är en nödvändig strategi om

syftet är att kunna generalisera resultatet till den population randomiseringen avsåg: samplet kan inte som ett resultat av bortfall förlora kopplingen till den population urvalet representerar. Det för med sig att varje observation ska analyseras så som den randomiserats och att forskaren designar och genomför studien på ett sådant sätt att avhopp minimeras och kompletta data maximeras. Därmed inkluderas utfallsmått för deltagare som av någon anledning hoppat av studien. Ibland måste man förmå deltagare som hoppat av att återkomma till uppföljningsmätningar. Men alla gör ändå inte det. Den enda återstående möjligheten är då att uppskatta vad utfallet skulle ha blivit om de hade svarat. Det är att imputera data.

Imputerade värden är kvalificerade gissningar eller uppskattningar som visserligen kan ha empiriskt stöd, men som är baserade på olika antaganden som ofta inte kan verifieras. Värden som saknas kan uppskattas med olika metoder, så kallade imputeringsmetoder. Den kompletta uppsättningen data efter en sådan process analyseras sedan med statistiska standardmetoder, som exempelvis variansanalys och regressionsanalys.

Det finns en mängd mer eller mindre sofistikerade metoder för att hantera bortfall, av vilka några presenteras här. Det är till hjälp att under presentationen tänka på att valet av metod beror på mekanismerna bakom bortfallet och de antaganden man gör om bortfallsmekanismerna. Dessa är inte möjliga att enkelt verifiera. På grund av detta dilemma kan man egentligen aldrig vara helt säker på att uppmätta behandlingseffekter är korrekta om de bygger på analys av imputerade data.

Om man har tillgång till fler imputeringsmetoder för en uppsättning data bör i vissa fall åtminstone två användas för att jämföra skillnader i effekter av de statistiska analyserna på imputerade data. En sådan procedur kallas ibland för sensitivitetsanalys.

> ***Imputering***
> (Av lat. i´mputo, tillgodoräkna.) Att ersätta ett saknat värde med ett verkligt med hjälp av statistik.

Typer av bortfall

Generellt sett finns två typer av bortfall: dels bortfall av enheter eller deltagare (enhetsbortfall eller individbortfall), dels bortfall av information på vissa individattribut (partiellt bortfall). Individbortfall uppkommer när data förloras eller inte samlas in på grund av att deltagare inte vill svara, avbryter medverkan, inte går att få tag i eller av andra skilda skäl är olämpliga eller ej godtagbara som deltagare (eng. *ineligible*). Partiellt bortfall uppkommer när en eller flera delfrågor i en enkät inte besvaras eller delar av en observation uteblir vid ett visst mättillfälle. Partiellt bortfall kan vara univariat (en fråga eller observation) eller multivariat (fler olika frågor eller observationer). Allmänt sett är potentiella systematiska fel (bias) i data en funktion av omfattningen av bortfallet; ju fler datavärden som saknas, desto högre är graden av systematiska fel i statistiken (Kline, 1998). Å andra sidan finns det ingen tydlig gräns för vad som kan betraktas som ett litet eller stort bortfall, men 1–5 procent har av somliga setts som ett litet bortfall och 15–20 procent som ett stort bortfall.

Bortfallsmönster och relaterade mekanismer

Imputeringsmetoder baseras på antaganden om att bortfallsmönstret är slumpmässigt eller systematiskt. Olika mekanismer bakom inkompletta data kan avgränsas genom att på ett probabilistiskt sätt villkora deras relation till kovariater vid studiens start. Tre mönster finns: fullständigt slumpmässigt bortfall, slumpmässigt bortfall eller inte slumpmässigt bortfall (Little & Rubin, 2002; 1987).

Fullständigt slumpmässigt bortfall

Om deltagare med värden som fallit bort och deltagare med kompletta värden antas vara två slumpmässiga urval från källpopulationen, kan bortfallet sägas vara fullständigt slumpmässigt (eng. *missing completely at random*, *MCAR*). Det betyder att det inte finns

något samband mellan värden på en variabel och vilket som helst värde som en annan variabel kan anta (inklusive kovariater och utfallsmått) och oberoende av om värdet är giltigt eller betecknar ett bortfall i sig självt. Bortfallet beror alltså inte på vare sig observerade eller förlorade data. Exempel på detta är när fel värde registrerats av ett trasigt instrument eller när data för en deltagare saknas för att frågeformuläret kommit bort.

Det finns ingen metod för att avgöra om bortfallsmekanismen verkligen är fullständigt slumpmässig, men det går att sluta sig till att den *inte* är det genom att jämföra gruppen deltagare som saknar värden med gruppen som har kompletta värden enligt definitionen ovan. Om det finns en skillnad mellan grupperna kan man vara säker på att bortfallet inte är fullständigt slumpmässigt, och om det inte finns någon skillnad är det fortfarande möjligt att skillnader finns i andra faktorer än dem som mätts i studien. Om exempelvis deltagare som saknar värden på en variabel X jämförs med deltagare som har kompletta värden angående någon egenskap vid studiens start (t.ex. socioekonomisk status) och det finns en signifikant skillnad är bortfallet inte fullständigt slumpmässigt. Om det inte finns någon signifikant skillnad kan man inte sluta sig till att bortfallet är fullständigt slumpmässigt eftersom det inte går att bevisa att skillnader beror på någon annan faktor som inte mätts i studien.

MCAR är ett starkt antagande om slumpmässighet i datamaterialet som sällan kan verifieras. Ibland antas MCAR felaktigt gälla när deltagare som fallit bort inte inkluderas i den slutgiltiga analysen. Det skulle innebära att uppföljningsdata för en viss behandlingsgrupp antingen observerats eller förlorats slumpmässigt och att bortfall inte medför mätfel i observationerna från den gruppen, oavsett omfattningen av bortfall i andra grupper.

Typer av bortfall

Fullständigt slumpmässigt bortfall (missing completely at random, MCAR) på en variabel råder om det inte finns något samband mellan saknade värden på en variabel och vilket som helst värde som en annan kan anta (inklusive kovariater och utfallsmått) och oberoende av om värdet är giltigt eller betecknar ett bortfall i sig självt.

Slumpmässigt bortfall (missing at random, MAR) då bortfall är enbart beroende av observerade data.

Inte slumpmässigt bortfall (missing not at random, MNAR) då saknade data är relaterade till deltagares sanna status på den variabeln.

Slumpmässigt bortfall

Ett svagare antagande angående bortfall är att det är slumpmässigt (eng. *missing at random, MAR*). Detta är ekvivalent med att säga att sannolikheten för bortfall enbart är beroende av observerade data (eng. *ignorability assumption*). Om data saknas eller inte saknas för en variabel är med andra ord inte relaterat till deltagares sanna status på den variabeln utan kan förklaras av andra variabler i datamaterialet (Hair, Anderson, Tatham & Black, 1998; Kline, 1998). Ett exempel på detta är om kvinnor i vissa länder är mindre benägna att uppge sin vikt än vad män är. Sannolikheten att ge den informationen är då inte en funktion av vikt, utan av kön.

Om MAR antas gälla kan imputeringen villkoras på observerade data eftersom sannolikheten att data saknas i en variabel inte beror på variabeln, utan på en annan variabels värden. Observera att deltagare med slumpmässigt bortfall inte är en slumpmässig undergrupp av studiepopulationen.

Icke slumpmässigt bortfall

Bortfall som inte kan ignoreras förekommer när bortfallet inte är slumpmässigt (eng. *missing not at random, MNAR*). Detta inträffar när förekomsten eller icke-förekomsten av bortfall på en variabel är relaterad till deltagares sanna status på den variabeln, till ex-

empel att de personerna med högst alkoholkonsumtion också har högst sannolikhet att hoppa av en studie. Bakom bortfall som inte kan ignoreras finns systematiska, inte slumpmässiga, faktorer som inte är uppenbara, observerade eller mätbara på annat vis (Little & Rubin, 2002). Deltagare kan exempelvis hoppa av en studie på grund av plötslig (och oförutsedd) försämring av hälsotillståndet eller överviktiga deltagare kan vägra att uppge sin vikt och bortfallet på variabeln vikt beror då på viktvariabelns sanna värden. Det finns imputeringsmetoder för MNAR som specificerar en modell för bortfallsmekanismen. Ofta är det dock en svår och komplex procedur som involverar områdesexpertis och andra statistiska antaganden som är svåra att verifiera. Den typen av metoder ligger utanför ramen för detta kapitel.

Identifikation av bortfallsmönster

Det finns inga metoder för att bevisa att bortfall verkligen är fullständigt slumpmässigt, MCAR. Däremot kan man jämföra undergruppen individer där bortfall finns med undergruppen individer som har valida värden och få information om att bortfallet inte är MCAR. Genom att dela samplet i två delar; en del som består av individer med valida värden på en variabel X och en annan del som består av individer med bortfall på samma variabel, och förslagsvis göra ett oberoende t-test av medelvärdesskillnaderna mellan dessa två grupper avseende någon annan variabel Y, kan man avgöra om det finns en statistiskt meningsfull skillnad i variabeln Y mellan de två grupperna. En sådan skillnad betyder att det finns en koppling mellan bortfallet på X och de observerade värdena på Y, vilket falsifierar antagandet om fullständig slumpmässighet. Två reservationer gäller dock:

1. Om många jämförelser görs ökar sannolikheten för en statistiskt signifikant skillnad. När massignifikanstestningar görs i en statistisk analys brukar någon typ av korrektion användas (t.ex. Bonferroni). I det här sammanhanget betyder det att en skillnad som

tyder på att bortfallet inte är MCAR med avseende på någon bakgrundsvariabel kan försvinna, särskilt om korrektionsfaktorn är konservativ. Mindre konservativa metoder, som Holms procedur, ökar sannolikheten för att signifikanta skillnader kvarstår.
2. Om samplet är stort kan även små skillnader bli signifikanta.

I stället för t-test kan punktbiseriala korrelationer beräknas mellan en variabel Z som indikerar bortfall i variabel X och en variabel Y av vilken bortfallet i X antas vara beroende. Höga signifikanta korrelationer mellan Z, en dikotom variabel som anger bortfall i X, och Y tyder på låg slumpmässighet, vilket överensstämmer med att data inte är MCAR men möjligtvis MAR, även om den undre gränsen för vad som kan betraktas som en hög korrelation är inte entydigt definierad (Huisman, 1999).

Metoder för imputering

Det finns flera olika imputeringsmetoder för att hantera bortfall, både för utfallsvariabler och för andra variabler som samlats in i en studie. En del av dessa betraktas som ad hoc och grova medan andra är mer sofistikerade. Oavsett detta reflekterar valet av en metod de antaganden man är beredd att göra angående typ av tillgängliga data, eventuellt bortfall, syfte med analysen och resultatets grad av trovärdighet. Dessa antaganden går inte alltid att verifiera, med följden att det inte finns någon garanti för resultatens validitet. Inga åtgärder kan fullt ut kompensera för problem orsakade av bortfall, men vissa åtgärder kan minimera problemen om hänsyn tas till osäkerheten som finns inbyggd i att imputera uteblivna data, exempelvis genom att regelbundet använda sensitivitetsanalys eller fler imputeringsmetoder. Imputering sker aldrig i ett vakuum utan är kontextberoende. Det ska framgå i dataanalyserna vilken procedur och vilka antaganden som gjorts samt alla omständigheter kring identifiering och hantering av bortfall. Fortsättningsvis presenteras bara metoder som är tillämpliga under antaganden om MAR

och MCAR. Om det inte uttryckligen framgår om imputeringen ska göras på hela uppsättningen data eller separat för de olika behandlingsgrupperna förutsätts det att sådana beslut fattas av forskaren på basis av kontext, design och planerade statistiska analyser.

Metoder för att flytta fram värden

Ett sätt att imputera bortfall och som har använts frekvent innebär att det senast observerade värdet flyttas fram (eng. *last observation carried forward*, *LOCF*). I denna metod ersätts en individs förlorade värden på en variabel av intresse med variabelns sista valida värde, oavsett tid mellan dessa två mättillfällen. Det sista mätvärdet antas vara en icke systematisk snedvriden uppskattning av individens sanna värde; ett antagande som inte går att verifiera.

En varning är på sin plats när LOCF används på grund av en del medföljande problem som har identifierats för vissa studiedesigner och särskilt om bortfallsmönstren är olika i studiegrupperna. Exempelvis kan en eventuell behandlingseffekt försvagas om utfallsmåttet utgörs av ett tidsberoende tillstånd som utvecklas eller försämras med tiden. Motsatsen kan inträffa om effekten av en behandling är tidsbegränsad, då behandlingseffekten i stället förstärks. Dessutom kan medelvärden och kovariansstruktur snedvridas i longitudinella studier med upprepad mätning och inomindividsvariansen (d.v.s. variansen för samma individ vid olika mättillfällen) kan spädas ut (Verbeke m.fl., 1997). På grund av dessa problem rekommenderas inte LOCF om det inte finns tydliga argument.

En mer konservativ variant av denna metod är att flytta fram baslinjemätpunktens värden (eng. *baseline carried forward, BCF*), att bortfall imputeras med baslinjevärden även om det finns data från mellanliggande mättillfällen. Det antas med andra ord att ingenting har hänt med utfallsvariabeln under studieperioden, vilket kan vara svårt att försvara om inte sensitivitetsanalys eller andra imputeringsmetoder också används.

Ytterligare en variant av denna metod tar tillvara alla valida data före det värde som ska imputeras (eng. *previous row mean/*

median, *PRM*). PRM ersätter det värde som saknas på en variabel med medelvärdet (eller medianen) av en individs alla valida värden på variabeln före det mättillfälle som saknar information. Det är viktigt att komma ihåg att extrema värden bland dessa kan snedvrida medelvärdet.

Metoder för omkringliggande värden

Imputeringsmetoden senast och nästa (eng. *last and next*) kan användas för att imputera bortfall vid ett mättillfälle mellan två observerade mättillfällen, ett före och ett efter. Mättillfället som fallit bort ersätts med genomsnittet av individens senaste och nästkommande valida information. Metoden kan givetvis inte användas för att imputera det sista mättillfället, men den kringgår några av problemen med LOCF för mellanliggande mätpunkter, särskilt om intervallet mellan den senaste mätpunkten och nästa mätpunkt är litet.

En variant på denna metod, som tar tillvara all valid information på en variabel när ett förlorat värde imputeras, är RM (”row mean/median”). Här beräknas medelvärdet av alla mätpunkters valida värden för en individ, vilket ersätter det förlorade värdet.

Relaterade metoder

I en metod ersätts ett förlorat värde på en variabel med ett observerat värde från ett mättillfälle efter det som fallit bort (eng. *next observation carried backward, NOCB*). Fler varianter som speglar metoderna att flytta fram värden går att föreställa sig.

Worst case method

En enkel men konservativ strategi som drar nytta av den allmänna riktningen i förändringen i behandlingsgrupper kallas sämsta alternativet (eng. *worst case, WC*). I den metoden imputeras ett förlorat värde i gruppen som fått insats (eller den grupp som förbättrats mest) med det sämsta utfallet (eller ett misslyckat fall) och i kontrollgruppen (eller den grupp som förbättrats minst) ersätts ett förlorat värde med det bästa utfallet (eller ett lyckat fall). Den strate-

gin blir extra konservativ om avhopp är vanligare i den grupp som fått insats (eller den grupp som förbättrats mest), eftersom behandlingseffekten då kan försvagas avsevärt.

Komplikationer kan också uppkomma om utfallet beräknas i antal (eller frekvens av en händelse) såsom antal fängelsedomar den senaste månaden, det vill säga det sämsta alternativet kan inte ha en tvetydig definition. Om man inte känner till gränsen för det sämsta tänkbara fallet kan det sämsta värdet tas från deltagarna med kompletta data. Nackdelen med det är att avhoppare inte sällan utgör en undergrupp med ännu sämre utfall än vad som rapporteras från deltagare med kompletta data.

Metoder som använder variablers medelvärde, median eller typvärde

En annan enkel strategi är att ersätta bortfall med medelvärdet eller medianen för den aktuella variabeln. Om data inte kan betraktas vara på kontinuerlig skala kan medianvärde användas i stället. Den här metoden utgår från att data är MCAR. Den bibehåller variabelns medelvärde (median eller typvärde), men har bland annat bieffekten att reducera den imputerade variabelns varians liksom korrelationer med andra variabler, särskilt när bortfallet är stort (Tabachnick & Fidell, 2001). Den är mer lämpad för studier där omständigheter såväl som deltagare inte är för heterogena.

Mer utarbetade strategier utgår från idén med imputering av medelvärden men villkorar att det imputerade värdet ska hämtas från den grupp individen med bortfallsvärden tillhör. Exempelvis ersätts det saknade värdet för en deltagare i behandlingsgruppen med medelvärdet från deltagare i behandlingsgruppen med kompletta data, medan en annan deltagares saknade värde i säg kontrollgruppen ersätts med medelvärdet från dem med kompletta data i kontrollgruppen. Ibland kan kovariater användas för att skapa ännu mer finkorniga klassificeringar, där medelvärdet i varje sådan klass får ersätta bortfall i en variabel för deltagare i den klassen.

Imputering med regression

Icke-stokastisk regression (eng. *non-stochastic regression, N-SR*) är en ansats i vilken förlorade data ersätts med imputerade värden predicerade i en regressionsmodell som innehåller relevanta prediktorer. Data från deltagare som saknar bortfall används då för att skapa en regressionsekvation där en variabel av intresse (t.ex. ett utfallsmått) kan prediceras med hjälp av andra variabler som antas orsaka eller korrelera med bortfallet på variabeln i fråga (t.ex. kovariater vid studiens start). Regressionsekvationen som blir resultatet används sedan för att imputera värden som fallit bort på den variabeln.

Som prediktorer i regressionsmodellen har några förespråkat att använda kovariater med starkast samband (Afifi & Elashoff, 1969; Raymond & Robert, 1987), medan andra har förespråkat variabler som är relaterade till bortfallet. Det har även argumenterats för att använda varje tillgänglig variabel i prediktionsmodellen (Little & Rubin, 2002; Allison, 2002). En annan synpunkt är att höga interkorrelationer (r = 0,2–0,3 eller högre) kan tyda på att denna metod är mer adekvat än andra metoder för en viss uppsättning data (Albridge, Standfish & Fries, 1988; Kaufman, 1988; Donner, 1982: Gleason & Staelin, 1975; Acock, 1997).

Även denna metod har sina begränsningar. Korrelationer och regressionskoefficienter kan överskattas och den aktuella variabelns varians beräknad på de imputerade fallen kan underskattas (Allison, 2002; Hair m.fl., 1998). Det senare är en konsekvens av att alla individer med samma värde på prediktorvariablerna kommer att få samma imputerade värde för det värde som fallit bort.

Imputering med stokastisk regression (SR) löser problemet med *underskattad* varians genom att addera en slumpkomponent (brus) till de predicerade värdena för att på så sätt göra varianserna av de imputerade värdena respektive de observerade värdena lika.

Hot deck (nearest neighbour method)

I metoden ”hot deck” ersätts förlorade värden i ett inkomplett fall med värden tagna från det fall som ”ligger närmast”. Därför kallas

metoden ibland ”närmaste grannen” (eng. *nearest neighbour method*). Den går ut på att en individ med bortfall på en variabel jämförs med den individ med fullständigt värde som är mest lik; ett komplett fall vars observerade värde på variabeln ifråga används för att imputera det saknade värdet. Den kritiska likheten som definierar en granne beräknas med hjälp av de variabler som inte behöver imputeras.

Det finns några val som måste göras när den här metoden används och resultatet kan variera beroende på dessa val. Undersökaren måste välja vilken typ av likhetsmått som ska användas. För kontinuerliga mått används ofta det euklidiska avståndet. Ibland kan det mest lika fallet (givare) väljas oavsett hur lika eller olika det är i förhållande till det fall som ska imputeras (mottagare). Detta inträffar när man inte kan eller vill definiera någon gräns i förväg för hur lika eller olika två individer ska vara för att bli betraktade som grannar utan man tar den närmaste grannens värde. I andra fall kan ett visst avstånd som inte får överskridas mellan givare och mottagare väljas. Om det senare alternativet väljs måste ett gränsvärde specificeras. Den här typen av imputering kan förbättras genom att dela samplet i olika klasser utifrån någon relevant kategori innan imputeringen görs.

Hot deck-metoder är inte beroende av antaganden om fördelningar och fungerar bra även i skeva fördelningar, och de imputerade värdena får samma form på fördelningen som de observerade data (Rubin, 1987). Se även Lessler och Kalsbeek (1992) och Reilly (1993).

Multipel imputering

Multipel imputering (eng. *multiple imputation, MI*) representerar en mer sofistikerad ansats att modellera och hantera bortfall och betraktas av många i dag som ”state of the art” inom området. Den centrala idén i de metoder som hittills presenterats om att ersätta varje uteblivet värde med ett värde draget från en uppskattning av variabelns distribution har här vidareutvecklats. I de metoder som nämnts tidigare tas det dock inte hänsyn till osäkerheten som finns i uppskattningen. Det gör däremot MI genom att ersätta ett förlorat

värde med ett antal värden som vart och ett är draget från lika många olika uppskattningar. Resultatet blir då inte *en* imputerad uppsättning data (*en* datamatris) utan ett antal uppsättningar imputerade dataset (vanligtvis 5–10 datamatriser), så kallade multipelt imputerade data. Efterföljande analyser kan då göras separat för varje uppsättning data och resultaten kan sedan kombineras för att få en samlad uppskattning av en behandlingseffekt och relaterad spridning.

I varje imputerat dataset (ofta fem till tio) har varje variabel med bortfall imputerats med uppskattade värden utifrån andra relaterade variabler, typen av bortfall eller båda dessa källor, genom exempelvis stokastisk regressionsimputation. I den enklaste formen av MI görs statistiska standardanalyser för varje imputerat dataset och resultaten kombineras på ett statistiskt lämpligt sätt. Även om proceduren kan genomföras manuellt, finns det statistikprogram (t.ex. SAS) som gör den automatiskt och producerar resultaten relevanta för valda statistiska analyser (som medelvärden, regression, ANOVA). Regler för hur analysresultaten från de imputerade dataseten ska kombineras till enhetliga estimat finns i exempelvis Rubin (1987).

Utöver stokastisk regressionsteknik är andra ofta använda metoder inom MI ”maximum likelihood” (ML) och ”expectation minimization” (EM). Båda metoderna är grundade i statistisk teori. ML nyttjar den villkorade fördelningen av variabeln ifråga givet kompletta data på prediktorvariablerna. När de relevanta parametrarna har uppskattats kan värdena som fallit bort på variabeln ifråga också uppskattas. För en mer detaljerad beskrivning, se Desarbo, Green och Carroll (1986) och Lee och Chiu (1990).

Uppskattning av värden med ML kan utgöra ett första steg i en upprepningsprocess där även parametrarna i modellen uppskattas igen efter imputering. På basis av de nya parametrarna görs en ny uppskattning av bortfallets värden. Denna process är EM, som nämndes ovan, och den fortsätter tills konvergens uppnås.

Om bortfallet är slumpmässigt, MAR, (eller fullständigt slumpmässigt, MCAR) resulterar både singel och multipel imputering i icke systematiskt snedvridna uppskattningar av associationer i data.

Uppskattningen av spridningen är inte valid i singel imputering, men däremot i multipel imputering – i och med att osäkerheten i imputerade värden tas med i beräkningen. MI-modellen, till skillnad från singel imputering, tar hänsyn både till eftersträvad allmän variation i populationen och till osäkerheten som finns inbyggd i uppskattningar av bortfall utan att ge avkall på sambanden mellan huvudvariablerna och övriga variabler (se även Dempster, Laird & Rubin, 1977; Laird, 1988; Ruud, 1991; Little och Rubin, 1987; Hedderley & Wakeling, 1995). För mer detaljerad statistisk information om MI, se Rubin (1987, 1996), Little och Rubin (2002) och Schafer (1997).

Andra metoder för att hantera bortfall

Metoden att bara använda kompletta fall (eng. *complete case analysis, CC,*) eller ”listwise deletion” imputerar inte bortfall och är därför ingen imputeringsmetod. Det som utmärker CC är att enbart fall med valida värden på alla variabler inkluderas i analysen (alla deltagare med bortfall exkluderas). Det kan innebära att en person med kompletta värden för nio mått men som saknar ett blir exkluderad. I många statistikprogram är CC standardinställning för analyser.

CC är acceptabelt att använda i reguljära analyser där modellen inte innehåller för många variabler som leder till att många deltagare förloras (omfattande förlust av information) och där det inte finns systematiska skillnader mellan kompletta och inkompletta fall. Är det inte så, minskar analysernas statistiska power och signifikanta systematiska fel kan möjligen förekomma.

Den här metoden är inte kompatibel med randomiserade kontrollerade studier eller ITT-principer om inte data är MCAR, ett antagande om bortfall som inte kan bevisas. Om data är MAR, kan analyser som använt CC bli missvisande. En CC-analys som har justerats för kovariater är dock rimlig under den inte lika begränsade förutsättningen kovariatberoende fullständig slumpmässighet.

Allmänna rekommendationer

Ett antal olika metoder för imputering har beskrivits. Valet av metod för imputering förutsätter vissa antaganden om data och bortfall som är förknippade med metoden, att omfattningen av bortfallet är hanterbart och följaktligen att resultaten är trovärdiga. Randomiserade kontrollerade studier är särskilt känsliga för systematiska fel i data på grund av bortfall, fel som kan påverka uppskattningen av behandlingseffekten, jämförbarheten hos behandlingsgrupperna liksom hur väl samplet representerar målgruppen.

Det är viktigt att komma ihåg att imputering är en form av mer eller mindre kvalificerade gissningar. För att välja rätt alternativ är det avgörande att känna till variablernas domän och deras fördelning liksom omständigheter kring datainsamlingen och genomförandet av studien. Imputering kan då vara en hjälp att minska de negativa konsekvenserna av bortfall. Men imputerade värden är inte sanna värden och kan inte heller behandlas som sådana. I första hand ska man planera och genomföra studien så att bortfall minimeras och helst undviks helt.

En nära relaterad fråga handlar om omfattningen av bortfallet i studien. Det finns inga generella regler men ett antal rekommendationer. Ett är dock säkert, och det är att ju mer bortfall, desto mindre pålitliga blir imputationerna. Metoder för singel imputering fungerar i allmänhet bra när omfattningen av bortfallet är mindre än fem procent och ”hot deck” eller regression när omfattningen är mellan fem och tio procent. Men eftersom MI är den mest robusta metoden är rekommendationen att använda den i alla tillämpliga fall.

Sensitivitetsanalys rekommenderas för att uppväga inverkan av en viss imputeringsmetod genom att jämföra slutresultaten av en behandlingseffekt med resultaten från en annan imputeringsmetod. Om resultaten är lika mellan olika antaganden och imputeringsmetoder är det ett tecken på att imputeringsresultaten är tillförlitliga.

Avslutningsvis kan extrema datavärden påverka imputeringsprocessen orimligt mycket beroende på vilken imputeringsmetod som

använts. Ett viktigt steg i varje imputeringsprocess är därför att identifiera extremvärden.

Fördjupningslitteratur

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549–576.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* London: Chapman and Hall.

Referenser

Acock, A. C. (1997). Working with missing values. *Family Science Review, 10*, 76–102.

Afifi, A. A. & Elashoff, R. M. (1969). Missing observations in multivariate statistics III, IV. *Journal of the American Statistical Association, 64*, 337–358.

Albridge, K. M., Standish, J. & Fries, J. F. (1988). Hierarchical time-oriented approaches to missing data inference. *Computers and Biomedical Research*, *21,* 349–366.

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum Likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, Series B, vol. 39, 1:1–38.

DeSarbo, W. S., Green, P. E. & Carroll, J. D. (1986). An alternating least squares procedure for estimating missing preference data in product concept testing. *Decision Sciences, 17*, 163–185.

Donner, A. (1982). The relative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing values. *American Statistician*, *36*, 378–381.

Gleason, T. C. & Staelin, R. (1975). A proposal for handling missing data. *Psychometrika, 40*, 229–252.

Graham J. W. & Hofer, S. M. (2000). Multiple imputation in multivariate research. I T. D. Little, K. U. Schnabel & J. Baumert (Red.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples* (s. 201–218). Mahwah, NJ: Lawrence Erlbaum.

Hair, J. F., Anderson, R. E., Tatham, R. L. & Black, W. C. (1998). *Multivariate data analysis with readings* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Hedderly, D. & Wakeling, I. (1995). A comparison of imputation techniques for internal preference mapping, using Monte Carlo simulation. *Food Quality and Preference, 6*, 281–297.

Hollis, S. & Campbell, F. (1999). What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ*, *319* (7211), 670–674.
Huisman, M. (1999). *Item nonresponse: Occurrence, causes, and imputation of missing answers to test items*. Leiden, Nederländerna: DSWO Press.
Kaufman, C. J. (1988). The application of logical imputation to household measurement. *Journal of the Market Research Society, 30*, 453–466.
Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York: Guilford.
Laird, N. M. (1988). Missing data in longitudinal studies. *Statistics in Medicine, 7*, 305–315.
Lee, S. Y. & Chiu, Y. M. (1990). Analysis of multivariate polychoric correlation models with incomplete data. *British Journal of Mathematical and Statistical Psychology, 43*, 145–154.
Lessler, J. T. & Kalsbeek, W. D. (1992). *Nonsampling Errors in Surveys*. New York: Wiley.
Little, R. J. A. & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley and Sons.
Little, R. J. A. & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York: Wiley.
Peugh, J. L. & Enders, C. K. (2004), "Missing data in educational research: A review of reporting practices and suggestions for improvement." *Review of Educational Research*, *74*, 525–556.
Raymond, M. R. & Roberts, D. M. (1987). A comparison of methods for treating incomplete data in selection research. *Educational and Psychological Measurement, 47*, 13–26.
Reilly, M. (1993). Data analysis using hot deck multiple imputation. *The statistician*, *42*, 307–13.
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley and Sons.
Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*, 473–489.
Ruud, P. A. (1991). Extensions of estimation methods using the EM algorithm, *Journal of Econometrics, 49*, 305–341.
Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman and Hall.
Schafer, J. L. & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147–177.
Tabachnick, B. G. & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston: Allyn and Bacon.
Veberke, G., Molenberghs, G., Bijnens, L. & Shaw, D. (1997). *Linear mixed models in Practice*. New York: Springer.

Kapitel 12 | Laura Ferrer-Wreder & James Jaccard

Förberedande datagranskning[1]

Detta kapitel handlar om vikten av förberedande analyser (eng. *data screening*) av insamlade data i effektutvärderingar, något som forskare ofta förbiser (Jaccard & Guilamo-Ramos, 2002). Dessa analyser är viktiga, eftersom de säkerställer datas integritet, att datamaterialet ger en rättvis bild av de data som samlats in. Generellt sett bör sådana analyser alltid göras före huvudanalyserna om resultatet ska gå att lita på. I vissa fall kan resultaten från datagranskningen även användas i de slutgiltiga utfallsanalyserna. En positiv bieffekt av en omsorgsfull datagranskning är att forskaren blir bekant med data. När det blir dags för tolkning kan denna förtrogenhet vara till hjälp vid klivet från resultat till genomtänkt tolkning. Trots att vissa statistiker har ifrågasatt den praktiska nyttan med vanligt förekommande datagranskningstest (t.ex. Wilcox, 2002), finns det bara fördelar med att få en bild av de data man ska arbeta med. Det går åt mycket tid, planering och energi för att utvärdera en intervention. I detta arbete utgör en förberedande datagranskning en viktig del.

Kapitlet är inriktat på behoven i effektutvärderingar där data vanligtvis analyseras med hjälp av variansanalys (ANOVA), kovariansanalys (ANCOVA) eller multipel regressionsanalys. Dessa ana-

1 Kapitlet har översatts från engelska till svenska av Åsa Kling, institutionen för psykologi, Uppsala universitet.

lysstrategier är känsliga för extremvärden (eng. *outliers*) och snedfördelning av observationerna (eng. *skewness*). Kapitlet begränsas därför till identifiering och hantering av:

- extremvärden
- bristande normalfördelning
- bristande homogen varians
- beroende observationer.

Huvudsyftet med en datagranskning är att identifiera ogiltiga värden, extremvärden samt att få överblick över fördelningen av variabler i en uppsättning data. Det kan både undersökas inom (univariat) och över (multivariat) variabler. Dessa två huvudsyften kan vara beroende av varandra, som när data inte är normalfördelade på grund av extrema värden (Wilcox, 2002).

De datagranskningsstrategier som beskrivs i kapitlet har behandlats utförligt i andra sammanhang, inklusive teori och beräkningsformler (t.ex. Maxwell & Delaney, 2004). Därför är kapitlet utformat som en kortfattad, selektiv översikt av den omfattande och ofta tekniska referenslitteraturen på området. Först presenteras metoder för att identifiera hot mot de vanligaste statistiska analyserna i effektutvärdering. Därefter beskrivs några metoder för att hantera dessa hot.

Det är viktigt att dokumentera det som görs i den förberedande datagranskningen så att man vid behov kan gå tillbaka och kontrollera.

För samtliga analyser som beskrivs här bör datagranskningen göras separat för interventions- respektive kontrollgrupp, liksom för varje mättillfälle.

Initiala analyser[2]

Inmatnings-, kodnings- och scanningsfel

Oberoende av om inmatning av data sker manuellt eller via elektronisk skanning, är det viktigt att kontrollera potentiella inmatnings-,

2 Se även kapitel 9.

kodnings- eller scanningsfel. Det ska vara en exakt överensstämmelse mellan den elektroniska datamatrisen och insamlade originaldata. I ett litet datamaterial kan det vara möjligt att kontrollera alla deltagare, men i ett medelstort eller stort datamaterial är det mest praktiskt att göra slumpmässiga stickprovskontroller av centrala variabler. Om helt elektronisk insamling och inmatning används behövs inte denna kontroll (t.ex. ALEXSA-systemet[3]; Ridenour, Clark & Cottler, 2009).

Oavsett hur observationerna samlats in och matats in i en datamatris rekommenderas att man skapar frekvenstabeller av alla variabler för att identifiera värden som ligger utanför den skala som använts i studien. Om svarsalternativen i en skala exempelvis varierar mellan 1 och 5 och siffran 99 finns i frekvenstabellen är 99 ett värde som ligger utanför den skalan. Sådana omöjliga skalvärden kan identifieras i frekvenstabeller för alla typer av variabler, kontinuerliga såväl som kategoriska, och de bör spåras tillbaka till datakällan (t.ex. pappersenkäten) om de upptäcks. Värden som hamnar utanför skalan kan bero på missar vid datainmatningen eller uteblivna värden som inte har angetts som bortfall (Raykov & Marcoulides, 2008).

Inkonsekventa och icke-trovärdiga svar

I en del effektutvärderingar, beroende på typ av instrument och syftet med interventionen, är det relevant att kontrollera svarskonsistens och svar som inte är trovärdiga. Vissa skattningsinstrument innehåller exempelvis frågor om användning av en påhittad drog eller låter deltagarna skatta hur sanningsenligt de besvarat frågorna. Ett exempel är om två frågor använts för att studera elevers missbruk av lösningsmedel (sniffning): (1) om eleven sniffat någon gång i livet (svarsalternativ: Ja eller Nej); (2) hur många gånger eleven sniffat under de senaste fyra veckorna (svarsalternativ: 0, 1, 2, 3, 4, 5, 6, 7, 8 eller 9 eller fler gånger). Deltagare som gett inkonsekventa svar är de som svarat nej på den första frågan och något annat svar

3 Assessment of Liability and EXposure to Substance use and Antisocial behaviour.

än noll på den andra. Om det förekommer deltagare med inkonsekventa svar ska det rapporteras. Dessutom kan forskaren överväga fördelar och nackdelar med att behålla eller utesluta dessa deltagare ur analyserna genom att undersöka om de har en betydelse för slutresultatet.

Kontroll av svarskonsistens

Kontroll av svarskonsistens kan göras med hjälp av SPSS. Om en deltagare exempelvis har besvarat frågan om sniffning av lösningsmedel har förekommit någon gång med att det inte skett (0 = Nej) och sedan på frågan om sniffning förekommit under de senaste fyra veckorna svarat att det förekommit (1 = Jag har sniffat lösningsmedel en gång) kan svarsmönstret för de två variablerna identifieras som inkonsekvent genom en ny variabel "Sniff_Konsist" (= 2). I funktionen "omkoda syntaxen" ("recode syntax") instrueras SPSS att ge värdet 1 till alla deltagare som inte har det angivna svarsmönstret för de här två variablerna.

Steg 1: Skapa en ny variabel (på sidan med variabelöversikten) kallad Sniff_Konsist och specificera etiketten ("value label") 1 = konsekventa svar och 2 = inkonsekventa svar.

Steg 2: Skriv och utför ("execute") syntaxen så som den skrivits nedan. GE 1 står för "greater than or equal" ("större än eller lika med"): IF (Variabl _ Sniff_Någongång = 0 AND Variable_Sniff_Nyligen GE 1) Sniff_Konsist = 2.
RECODE Sniff_Konsist (SYSMIS = 1).
EXECUTE.

Extremvärden

Eftersom kontinuerliga variabler ofta används för de centrala frågeställningarna i effektutvärderingar, begränsas diskussionen till dessa variabler.[4] Extremvärden (eng. *outliers*) är värden som påtagligt avviker från övriga värden i en fördelning av data (Barnett & Lewis, 1994). Ett problem med dem är att de kan ha en orimligt stor

4 Se Zijlstra, van der Ark och Sijtsma (2007) för identifiering av extrema värden i kategoriska variabler.

inverkan på medelvärden och standardavvikelser och till följd av det försvåra för forskaren att upptäcka medelvärdesskillnader (Wilcox, 2002) vilket kan få konsekvenser för typ I- och typ II-fel (Tabachnick & Fidell, 2007). Om det finns extremvärden uppstår frågan om medelvärdet av en viss variabel verkligen är representativt för populationens medelvärde (Jaccard & Guilamo-Ramos, 2002). Ett annat bekymmer med extremvärden är att de kan ha en orimlig inverkan på ett datasets normalfördelning (Wilcox, 2002), liksom att de kan begränsa generaliserbarheten av studiens resultat (Tabachnick & Fidell, 2007).

Univariata extremvärden

Med univariata extremvärden avses extremvärden i en variabel. Ett univariat extremvärde är ett svar som inte överensstämmer med många av de andra deltagarnas svar på samma fråga. Om en deltagare i en studie exempelvis uppger en årslön på tio miljoner kronor, medan alla andra deltagare uppger årslöner på mellan tvåhundratusen och femhundratusen kronor, har den deltagaren ett extremt värde på variabeln årslön.

En vanlig strategi för att upptäcka univariata extremvärden är att kontrollera deltagarnas *z*-värden (standardiserade värden) för varje kontinuerlig variabel i studien. Det finns dock olika uppfattningar om kriteriet för vad som är ett extremvärde när *z*-värden används och detta kriterium kan även vara relaterat till storleken på urvalet (jfr Meyers, Gamst & Guarino, 2006; Tabachnick & Fidell, 2007). Ett kriterium som angetts för ett univariat extremvärde är ett absolut *z*-värde större än tre (Raykov & Marcoulides, 2008).

Att undersöka z-värden kan vara ett första steg i att bekanta sig med variablerna i en uppsättning data. Men att använda z-värden för att upptäcka univariata extremvärden är problematiskt på grund av ”maskering” (Wilcox, 2003); att extremvärden inte upptäcks på grund av att metoden för att upptäcka extremvärden bygger på medelvärden och standardavvikelser som själva är missvisande på grund av extremvärden (Wilcox, 1998; 2002). Jaccard och Guilamo-Ramos

(2002) har gett ett exempel där de använde ett traditionellt gränsvärde på två standardavvikelser från medelvärdet för att identifiera univariata extremvärden:

> … antal dagar som ungdomar har använt droger den senaste månaden kan visa en förekomst av 0, 0, 0, 0, 1, 1, 1, 2, 2, 3, 20, 30. Medelvärdet är 5,0 och standardavvikelsen är 9,62. I denna serie förefaller siffrorna 20 och 30 vara extremvärden. Men 20 är bara 1,56 standardavvikelser från medelvärdet, så det markeras inte som ett extremvärde. (s. 288)

En mer användbar strategi för att upptäcka univariata extremvärden är att använda en ”box-plot” (Tukey, 1977). Det är en grafisk illustration av en variabels data och kan med fördel användas för att upptäcka univariata extremvärden (Barnett & Lewis, 1994; Wilcox, 2002). En box-plot visar percentiler och medianen för rangordnade värden. Lådan i box-plotten omfattar värden inom den interkvartila räckvidden (IQR) definierad som intervallet mellan den 75:e och den 25:e percentilen (d.v.s. övre och nedre kvartilen; Tukeys ”kritiska punkter” [eng. *hinges*]; Meyers m.fl., 2006). En box-plot innehåller även en linje som representerar medianen eller den punkt där värdena delas i två lika stora halvor, när värdena rangordnats. Linjer eller ”morrhår” (eng. *whiskers*) går ut från boxen till ett tvärstreck på var sida. Morrhåren och tvärstrecken visar värdenas räckvidd utanför IQR. Dessa värden är däremot inte tillräckligt udda för att betraktas som extrema (Wilcox, 2003).

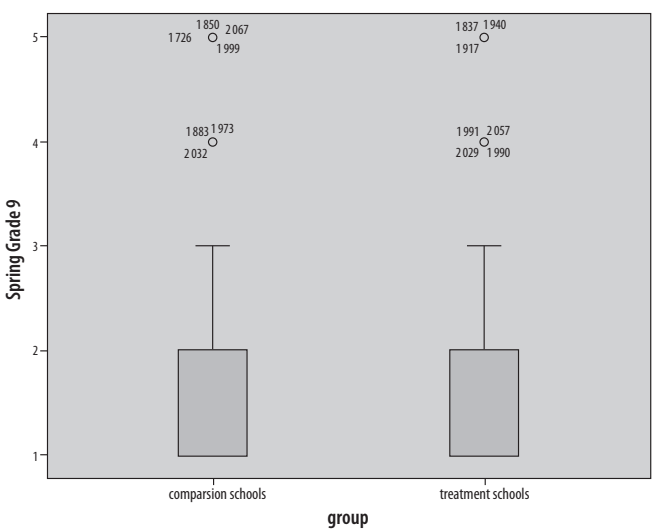
En box-plot i sin standardform har därmed begränsningar precis som de flesta diagnostiska verktyg. Exempelvis kan extremvärden maskeras om en fjärdedel eller fler av observationerna på en variabel är extremvärden, och användbarheten av box-plot kan variera beroende på stickprovsstorleken (Wilcox, 2003). I de flesta situationer är dock box-plot en användbar strategi för att upptäcka univariata extremvärden.

Exempel på SPSS syntax box-plot över en variabel med grupperade data (separat för interventions- och jämförelsegrupp)

En box-plot illustrerar svaren på frågan "Röker du?" med svarsalternativen 1 = Nej, har aldrig rökt; 2 = Nej, har bara smakat; 3 = Nej, har rökt men slutat; 4 = Ja, ibland men inte varje dag (typ feströkare); 5 = Ja, dagligen.

Steg i SPSS: Skriv och utför syntaxen så som den skrivits nedan (röker_v9 är den univariata skalan eller frågan av intresse; pilotskola är grupperingsvariabeln som skiljer interventionsgruppen från jämförelsegruppen).

```
EXAMINE VARIABLES=röker_v9 BY pilotskola
 /PLOT=BOX-PLOT
 /STATISTICS=NONE
 /NOTOTAL
 /ID=Lpnr.
EXECUTE.
```



Tolkning: Box-plot visar att medianen är ungefär 1 (vilket motsvarar "Nej, har aldrig rökt", strecket längst ner). Lådans längd visar den interkvartila räckvidden för båda grupperna (jämförelsegruppen till vänster och interventionsgruppen till höger). Siffrorna i box-plot är id-nummer för deltagare med värden som är univariata extremvärden (stjärnor är särskilt extremt avvikande värden och cirklar är extremvärden, båda betraktas som extremvärden så som de definierats i det här kapitlet).

Multivariata extremvärden

Multivariata extremvärden är extremvärden som berör flera variabler. En deltagare i en studie kan exempelvis ha ett specifikt svarsmönster på två variabler som är ovanligt och avviker från de övriga deltagarnas eller undergruppens svarsmönster. Deltagaren kanske har ett väldigt lågt värde på ett begåvningsmått och samtidigt extremt hög nivå av extraversion (utåtvändhet) i förhållande till resten av samplet. Den deltagaren identifieras då som extrem i en uppsättning variabler (d.v.s. ett multivariat extremvärde).

När multivariata extremvärden identifieras tittar man på deltagares svarsmönster eller värden på alla variabler av intresse i det så kallade multivariata rummet. Varje deltagare i en studie har en positionspunkt eller en uppsättning koordinater i det multivariata rummet. Varje sådan punkt representerar deltagarnas specifika kombination av värden på samtliga variabler i en uppsättning data (Tabachnick & Fidell, 2007, s. 74). Beroende på typ av data kommer många av dessa koordinater att bilda multivariata punktmoln eller svärmar (Rousseeuw & Zomern, 1990). Ett multivariat extremvärde är en punkt som avviker från de andra punkterna i molnet. Det kan vara punkter som ligger en bit bort från punktmolnet och/eller utanför det linjära mönster som punkterna i punktmolnet bildar (Rousseeuw & Zomern, 1990). Ett multivariat extremvärde har betydelse om det identifieras som ett ”hävstångsvärde” (eng. *leverage point*), det vill säga ligger väsentligt utanför punktmolnet. Ett hävstångsvärde ligger med andra ord *både* utanför punktmolnet *och* utanför (eller på tvärs med) det linjära mönstret som punktmolnet bildar (d.v.s. vad som i regressionsanalys avses med ett extremvärde; Rousseeuw & Zomern, 1990; Tabachnick & Fidell, 2007; Wilcox, 2003).

Multivariata extremvärden kan identifieras på olika sätt i enkel linjär och multipel regressionsanalys. Generellt sett går diagnostiken för multivariata extremvärden ut på att identifiera dem och bedöma vilken påverkan de har (d.v.s. om extremvärden har betydelse för de huvudsakliga utfallsanalyserna). Ofta jämför man då precisionen i resultaten av en regressionsanalys där extremvärden tagits

med i analysen mot resultaten av en analys där extremvärden inte är med. Analysen av multivariata extremvärden är beroende av vilka variabler som ska användas i utfallsanalysen.

En metod för att upptäcka multivariata extremvärden är Mahalanobisdistansen. Med dess hjälp beräknas det multivariata avståndet mellan en viss deltagares värden och genomsnittet av de övriga deltagarnas värden i en uppsättning data (Tabachnick & Fidell, 2007). Testet har nackdelar som inte gör det till ett idealiskt diagnostiskt verktyg, såsom att det inte fungerar bra när data inte är normalfördelade liksom att resultatet varierar beroende på urvalets storlek (Raykov & Marcoulides, 2008; Tabachnick & Fidell, 2007). En annan diagnostik som är matematiskt förknippad med Mahalanobisdistansen är ”hävstångsvärdet” (eng. *leverage value*). Dessa värden hjälper till att identifiera multivariata koordinater (och deltagare med dessa koordinater) som riskerar att förvrida testen av modellerna i en regressionsanalys (Everitt, 2002; Rousseeuw & Zomern, 1990; Yuan & Hayashi, 2010). Diagnostik med hävstångsvärden bör göras i kombination med andra multivariata diagnostiska strategier, som exempelvis standardiserade DfBeta-värden (se nedan).

Beräkningen av hävstångsvärden är oberoende av de utfallsanalyser som görs och motsvarande relationer som positionerar variabler i den typen av analyser. Hävstångsvärden kan skapas för varje deltagare genom att man gör en linjär regressionsanalys med relevanta variabler. Själva regressionen är inte det viktiga här, regressionen används som ett verktyg för att identifiera multivariata extremvärden.

Hävstångsvärden ger en bild av hur en deltagares multivariata koordinater ligger i förhållande till det multivariata punktmolnet (Rousseeuw & Zomern, 1990). Det är en indikation på betydelsen eller konsekvensen av ett multivariat extremvärde. Det bästa är om multivariata extremvärden undersöks med diagnostiska test som kan upptäcka om ett extremvärde är problematiskt eller inte (Yuan & Hayashi, 2010, s. 336). Problematiska hävstångsobservationer ligger långt utanför punktmolnet och går på tvärs med det linjära mönster som bildas av merparten av punkterna i punktmolnet (Rousseeuw

& Zomern, 1990; Tabachnick & Fidell, 2007). Se Rousseeuw och Zomern (1990) samt Yuan och Hayashi (2010) för fler exempel på skillnader mellan olika typer av hävstångsvärden (eller observationer) och extremvärden.

En strategi för identifiering av multivariata extremvärden är att beräkna och tolka betydelsen av hävstångsvärden kombinerat med standardiserade DfBeta-värden. Standardiserade DfBeta-värden ger en indikation på den standardiserade förändringen av en regressionskoefficient (betakoefficient) när en deltagare ingår i en multipel linjär regressionsanalys jämfört med när deltagaren utesluts ur analysen (Everitt, 2002). Hävstångsvärden och DfBeta-värden kan beräknas med multipel regressionsanalys, men kan också användas mer generellt för att identifiera multivariata extremvärden med betydelse för resultatet.

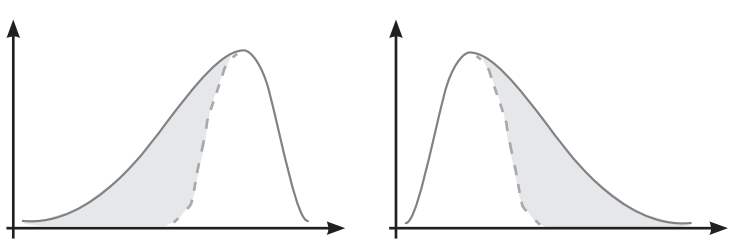
Standardiserade DfBeta-värden undersöks inom ramen för de specifika relationerna mellan de variabler som har definierats och som testas i utfallsanalyserna. Hur DfBeta-värden beräknas beror på vilka utfallsanalyser som görs och vilka hypoteser som finns. En studie som exempelvis prövar en modell för mediering av förändring genom intervention (d.v.s. undersöker förändringsteorin bakom en intervention) behöver använda flera regressionsekvationer eller path-koefficienter i modellen. En intervention förväntas t.ill exempel påverka användning av alkohol via en utvecklad förmåga att stå emot grupptryck och mer realistiska föreställningar om användning av droger. I den här typen av modell beräknas DfBeta-värden för varje regressionsekvation som ingår i analysmodellen. En av dessa ekvationer har prediktorerna förmåga att stå emot grupptryck respektive realistiska föreställningar om droger samt användning av alkohol som beroende utfallsvariabel. Tre DfBeta-värden beräknas då för varje deltagare, ett för interceptet, ett för prediktorn förmåga att stå emot grupptryck och ett för prediktorn realistiska föreställningar om droganvändning. Ett kriterium som indikerar möjliga multivariata extremvärden är DfBeta-värden som överstiger ett absolut värde av 1 (Jaccard, 2006).

Normalfördelning

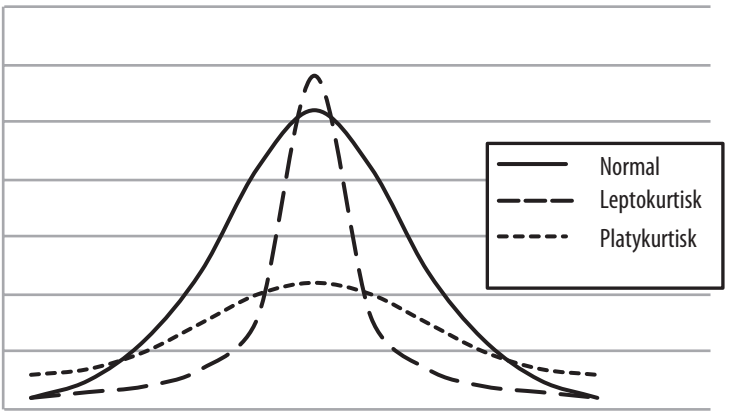
När man undersöker hur variablerna i en datauppsättning är fördelade är det fördelningens form man huvudsakligen intresserar sig för (DeCarlo, 1997). De flesta statistiska test som används i effektutvärderingar (t.ex. ANOVA, ANCOVA och multipel regression) förutsätter normalfördelade kontinuerliga variabler. Exakt hur dessa antaganden om normalfördelade observationer ser ut varierar beroende på statistiskt test och forskningsdesign. Normalfördelningsantagandet för ANOVA-statistik är exempelvis ”beroendevariabeln (t.ex. eftermätningen, måttet på skillnaden mellan mätningarna) är normalfördelad i populationen inom varje grupp (och beroende av förmätningen när ANCOVA används)” (Rausch, Maxwell & Kelley, 2003, s. 468).

Univariat normalfördelning

När man undersöker antagandet om univariat normalfördelning undersöker man om en kontinuerlig variabels form eller fördelning avviker från en klockformad normalkurva på ett sådant sätt att studiens statistiska styrka (eng. *power*) påverkas negativt (Wilcox, 2002). Kurvans skevhet (eng. *skewness*) och kurtos (eng. *kurtosis*) är två statistiska mått på antagandet om univariat normalfördelning. Vid skevhet är normalkurvan inte symmetrisk utan planar ut mer åt ena sidans ände (Tabachnick & Fidell, 2007). Några exempel på skevhet presenteras i figur 12:1.



Figur 12:1. Grafisk beskrivning av negativ och positiv skevhet.



Figur 12:2. Grafisk beskrivning av positiv och negativ (leptokurtisk och platykurtisk) kurtos.

En fördelning med positiv kurtos kallas även leptokurtisk med hög, smal topp kring medelvärdet samt tjocka svansar. Sannolikheten för extrema utfall är då hög jämfört med en normalfördelning (figur 12:2). En fördelning med negativ kurtos kallas även platykurtisk och har en tjock topp med smala eller inga svansar, vilket ger den en plattare form (DeCarlo, 1997, s. 292). En intressant detalj är att om en normalkurva läggs över fördelningar med positiv eller negativ kurtos ska bägge dessa olika fördelningar korsa normalkurvan till vänster och höger om medelvärdet.

Att ta fram statistik för skevhet och kurtos för varje kontinuerlig variabel kan utgöra första steget i processen att undersöka antagandet om normalfördelade data. Ett kriterium för att upptäcka problematisk skevhet och kurtos i en fördelning av en kontinuerlig variabel är absoluta kurtosvärden högre än två (Raykov & Marcoulides, 2008).

Exempel på kontroll av univariat skevhet ("skewness") och kurtos

Stegen i SPSS är "Analyze → Descriptive Statistics > Descriptives" → klicka för att flytta variablerna av intresse från "Variables" till vänster, ange "Förmätning Alkohol (FÖREALKOHOL) och Positiva Känslor (förepositiv)" till rutan på höger sida → under "Options", klicka på "Skew och Kurtosis → Okay".

Instruera SPSS att ta fram en tabell för skevhet (Skewness) och kurtos enligt:
DESCRIPTIVES VARIABLES = FÖREALKOHOL förepositiv
/STATISTICS = MEAN STDDEV MIN MAX KURTOSIS SKEWNESS.

	N	Skewness		Kurtosis	
	Statistic	Statistic	Std. Error	Statistic	Std. Error
FÖREALKOHOL	431	1,670	,118	2,814	,235
förepositiv	410	1,019	,121	,567	,240
Valid N (listwise)	405				

Tolkning: Indexet för "FÖREALKOHOL" överstiger ett absolut värde av 2 i jämförelsegruppen och därför avviker den variabeln från normalfördelningen i termer av kurtos.

Multivariat normalfördelning

Multivariat normalfördelning är antagandet om att varje variabel och alla linjära kombinationer av variablerna är normalfördelade (Tabachnick & Fidell, 2007, s. 78). Multivariat normalfördelning förutsätter med andra ord univariat normalfördelning, men univariat normalfördelning är ingen garanti för multivariat normalfördelning. Det är det där med "alla linjära kombinationer" som gör det svårt att fastställa multivariat normalfördelning på ett övertygande sätt, eftersom det skulle kräva att en avsevärd mängd potentiella kombinationer undersöktes (Raykov & Marcoulides, 2008). Om multivariat normalfördelning är svårfångat, kan man ändå få en uppfattning om potentiell multivariat normalfördelning genom flera strategier, bland annat genom att identifiera avvikelse från univariat normalfördelning i kombination med en omdömesgill användning av multivariata normalitetstest.

Ett multivariat test på normalfördelning är Mardias mått (Mardia, 1970; 1974; 1980) som ger ett index på multivariat skevhet och kurtos. Avvikelse från multivariat normalfördelning anges i de här måtten med kritiska värden över ett fastställt gränsvärde (d.v.s. om testen är signifikanta kan det finnas multivariat icke-normalitet). Ett exempel på hur multivariat normalfördelning kan testas är Mardias mått på multivariat kurtos. Måttet anger graden av oregelbundenhet i urvalsfördelningens svansar – det vill säga om ändarna i datafördelningen är lättare eller tyngre än vad som skulle förväntas i en multivariat normalfördelning (DeCarlo, 1997). De kurtosproblem som indikeras genom signifikanta värden på Mardias mått innebär att alltför många deltagare får multivariata koordinater som ligger långt från de övriga deltagarnas genomsnitt (d.v.s. tyngdpunkten – centroiden – av alla fall; DeCarlo, 1997, s. 298).

Mardias mått på skevhet och kurtos är begränsade i så måtto att de är mycket känsliga för mindre avvikelser från normalfördelningen som inte behöver vara av vikt (d.v.s. de ger ingen information om graden av avvikelse från normalfördelningen, bara att en avvikelse är sannolik). Dessutom kan de ge olika resultat beroende på urvalets storlek. Det kan vara värt att undersöka multivariat normalfördelning med Mardias mått, särskilt ihop med information från univariata normalfördelningstest. Men måtten måste tolkas med försiktighet. Mardias mått kan fås via makron (tilläggskommandon som kan hämtas via internet) för SPSS och SAS, och finns i program för strukturell ekvationsmodellering som AMOS.

Homogen varians

Ett antagande för att använda ANOVA och liknande statistiska test är att variansen i utfallsvariabeln ska vara homogen, det vill säga lika i alla grupper. Måttliga variansskillnader mellan grupperna spelar inte så stor roll om det rör sig om små stickprov, men vid större stickprov är det viktigt att de olika grupperna har samma varians.

Test av medelvärdesskillnader är inte särskilt motståndskraftiga,

robusta, mot överträdelser av antagandet om homogen varians, särskilt när jämförelsegrupperna är olika stora eller urvalet är litet (se Maxwell & Delaney, 2004; Wilcox, 2003; 2005). När antagandet om homogen varians inte uppfylls minskar den statistiska powern, även om observationerna är normalfördelade (Wilcox, 2003). *Levenes test* visar (via ett signifikanstest) om antagandet om lika varians håller genom en analys av residualerna mellan observationer och deras medelvärden (Everitt, 2002). Flera av den här sortens diagnostiska test är dock problematiska (Wilcox, 2002). En begränsning är problemet att vissa av testen är mest användbara när observationerna är normalfördelade, vilket de inte alltid är i praktiken (Parra-Frutos, 2009). Så detta är en del av datagranskningen som man behöver vara särskilt uppmärksam på.

Oberoende observationer

Generellt sett är F-testen i ANOVA inte särskilt robusta mot överträdelser av antagandet om oberoende data (Maxwell & Delaney, 2004). Om observationerna är relaterade till varandra på ett systematiskt sätt håller exempelvis inte antagandet om oberoende. Kenny och Judd (1986) har beskrivit oberoende i termer av villkorlig sannolikhet; varje observation inom en grupp är beroende av de andra observationerna i gruppen. Ett exempel är om skolklasser randomiseras till ett av två nya utbildningsprogram om hälsa. Utfallsmåttet är elevernas kunskap om hälsa i slutet av skolåret. Det är troligt att eleverna i en klass påverkar varandra. En driftig elev kanske får de andra eleverna i klassen mer engagerade så att de lär sig mer. I en sådan situation är elevernas värden inom en klass relaterade till varandra, vilket bryter mot antagandet om oberoende observationer. Ett annat exempel är om en utvärdering omfattar flera personer som fått behandling av samma terapeut.

Brott mot antagandet om oberoende är en fråga om experimentell design, som när forskaren beslutar att randomisera klassrum i stället för individer. I designer som randomiserar fristående delta-

gare till grupper som sedan jämförs är antagandet om oberoende rimligt. Men när observationer är relaterade till varandra på ett systematiskt sätt finns det statistiska metoder som tar hänsyn till beroendet i data (se kapitel 14).[5]

Metoder för att hantera problem som den förberedande datagranskningen identifierat

Den förberedande datagranskningen ger ett underlag för informerade beslut om eventuella lämpliga åtgärder. Om exempelvis extremvärden har upptäckts är den första åtgärden att gå tillbaka till källan och så långt det är möjligt kontrollera eventuella inmatnings-, kodnings- och scanningsfel och om extremvärdena möjligen har åstadkommits av deltagare som egentligen inte tillhör det avsedda samplet (Barnett & Lewis, 1994).

Hantering av extremvärden

Om det trots dessa kontroller finns ”sanna” extremvärden är det viktigt att undersöka om de deltagare som har univariata eller multivariata extremvärden har någonting gemensamt. Om de har det är den informationen viktig för utfallsanalyserna och den avslutande resultatdiskussionen. Efter att ha bekantat sig närmare med de deltagare som har extrema värden kan också den totala andelen sådana deltagare bestämmas i förhållande till deltagare i urvalet som saknar extremvärden.

Wilcox (1998; 2002) menar att exklusion av ”sanna” extremvärden skapar ett beroende i data som undergräver den statistiska teorin som många statistiska test bygger på. Inte minst blir tillämpning av standardfelet missvisande. Alla traditionella metoder för att dra generella slutsatser utifrån resultatet är inte heller lämpliga att använda om extremvärden exkluderats.

5 För mer information om klusterdata, se Judd, McClelland & Ryan (2009), Kenny & Judd (1986) samt Kenny, Kashy & Cook (2006).

Med dagens kunskap är det bästa sättet att hantera extremvärden att använda robusta estimatorer (Wilcox, 2003; 2005). Robusta estimatorer blir allt mer tillgängliga och vanligt förekommande. I SPSS funktion ”utforskaren” (”explore”) finns till exempel några robusta estimatorer. Inte minst är det viktigt att veta att det finns motsvarande robust statistik för många traditionella statistiska test, såsom de för jämförelser av medelvärden. De finns i datorprogrammet S-PLUS och i gratisprogrammet R (tillgängliga via internet).

Ett exempel på fördelar med att använda robusta estimatorer är trimmade medelvärden (eng. *trimmed means*) (Wilcox, 2002; 2003; 2005). Trimmade medelvärden används i robust statistik i stället för traditionellt beräknade medelvärden. I ett trimmat medelvärde har en bestämd andel av de högsta och lägsta värdena (20 procent) uteslutits (”trimmats bort”). De uteslutna värdena kan vara extremvärden eller inte. Ett aritmetiskt medelvärde beräknas på de återstående värdena och representerar det trimmade medelvärdet (Wilcox, 2002). Eftersom värdena först är rangordnade ingår alla insamlade värden inklusive extremvärden i det trimmade medelvärdet, men med fördelen att det inte påverkas av extremvärden.[6] Om trimmade medelvärden beräknats kan man inte sedan använda traditionella inferentiella statistiska test med dessa förändrade medelvärden (Wilcox, 2003). För trimmade medelvärden ska de robusta motsvarigheterna till de statistiska analyser man vill göra (t.ex. ANOVA, regressionsanalys) användas (Wilcox, 2003). Förutom att trimmade medelvärden tar hänsyn till extremvärden förbättrar de även måttliga avvikelser från normalfördelningen på grund av skevhet (Wilcox, 2002).

Om det inte är möjligt att använda robust statistik kan ett alternativ vara att avgöra om de identifierade extremvärdena verkligen påverkar de huvudsakliga utfallsanalyserna. Utfallsanalyser kan göras med eller utan extremvärden (Zijlstra m.fl., 2007). Om skillnaden

6 Se Wilcox (2003; 2005) för en introduktion till robust statistik och praktisk tillämpning av teknikerna med hjälp av datorprogrammen S-PLUS och R.

är av mindre betydelse när extremvärden ingår i analysen kan detta tas med i beskrivningen av resultaten tillsammans med proceduren för att identifiera extremvärden och resultatet av datagranskningsprocessen (t.ex. ”20 deltagare i jämförelsegruppen hade antingen univariata eller multivariata extrema värden i för- och/eller eftermätningen. Dessa deltagare representerade en procent av den totala jämförelsegruppen”). Om studiens resultat däremot skiljer sig signifikant när extremvärden ingått i analysen har Raykov och Marcolides (2008) föreslagit att båda resultaten med och utan extremvärden ska rapporteras. Dessutom bör det klargöras att generaliserbarheten för resultaten är begränsad när extremvärden exkluderats och bara gäller för individer som liknar dem i urvalet utan extrema värden.

Hantering av hot mot normalfördelning och homogen varians

Alternativa sätt att hantera problemen med data som inte är normalfördelade och har olika varians har komplikationer. En anledning är att de relevanta diagnostiska testen i vissa fall är problematiskt beroende av varandra; testen i sig kan ge begränsad information och kan oavsiktligt användas så att resultatet blir till begränsad nytta. När det gäller problematiskt ömsesidigt beroende är vissa metoder för diagnostik av variansers homogenitet bara användbara när data är normalfördelade (Parra-Frutos, 2009). Normalfördelning är med andra ord en förutsättning för vissa test. I andra fall kan hot mot homogen varians och normalfördelning identifieras. Men det är omfattningen av avvikelsen som är viktigast och forskaren måste avgöra om det är en betydelsefull avvikelse från normalfördelningen eller inte. Många av de diagnostiska test som finns kan bara säga om det existerar normalfördelning och homogen varians, de säger inte hur omfattande avvikelsen är. Om testen används felaktigt kan de diagnostiska testen dessutom påverkas orimligt mycket av urvalsstorlek, där både små och stora urval har sina särskilda problem (Wilcox, Charlin & Thompson, 1986; Wilcox, 2003). Utöver detta menar Wilcox (2002) att testen av olikheter i varians saknar ”förmåga att upptäcka fall där varianserna skiljer sig tillräckligt mycket

för att orsaka problem – även då data är normalfördelade" (s. 404). I princip lönar det sig att undersöka giltigheten av antagandena om normalfördelade data och homogen varians. I verkligheten måste sådana kontroller användas med försiktighet och med medvetenhet om de begränsningar som finns.

En strategi när den förberedande datagranskningen visar på hot mot normalfördelning och homogen varians är att använda "bootstrap"-metoden (Wilcox, 2003). Eftersom det finns många tekniker för bootstrap, beskrivs endast den generella innebörden av metoden. I korthet används bootstrap för att få bästa möjliga information om de data som finns tillgängliga. Metoden "omsamplar med återläggning" (Wilcox, 2002, s. 409). Upprepade sampel dras slumpmässigt från insamlade data. Varje gång en observation dras från det ursprungliga samplet återförs den dit och på det sättet ökar variationen i det nya samplet. Mått av olika slag, såsom medelvärde, kan sedan beräknas för alla nya sampel som dragits ur originalsamplet (Wilcox, 2003). Bootstraptekniker kan med gott resultat användas ihop med robusta estimatorer som trimmade medelvärden eller traditionellt beräknade estimatorer (Wilcox, 2003). Teknikerna finns i datorprogram som S-PLUS och R, samt i program för strukturell ekvationsmodellering som AMOS och Mplus. En detaljerad beskrivning av användning av bootstraptekniker med instruktioner steg för steg finns i Wilcox (2003).

Det finns ingen enighet om nyttan med att transformera data. Datatransformering för att exempelvis åtgärda skillnader i varians kan påverka normalfördelningen (Blaylock, Salathe & Green, 1980; Budescu & Applebaum, 1981; Doksum & Wong, 1983; Milligan, 1987; Wilcox, 1998). Datatransformering kan också försvåra tolkningen av resultaten och bidra till ofördelaktiga förändringar av sambanden mellan kovariater och utfallsvariabler, vilket kan öka risken för specifikationsfel och minska kovariaternas funktion av kontroll. Precis som i fallet med att exkludera extremvärden kan nackdelarna med transformering överstiga fördelarna.

Sammanfattning

Det här kapitlet behandlar vikten av förberedande datagranskning i effektutvärderingar. En sådan granskning bidrar till att säkerställa datas trovärdighet. I vissa fall kan resultaten från datagranskningen också ge information som behövs i utfallsanalyserna. En förberedande datagranskning kan innehålla följande steg:

- Kontrollera att data ger en rättvis bild av de data som samlats in (t.ex. skapa frekvenstabeller över alla variabler för att kontrollera värden utanför skalan etc.).
- Hantera eventuellt bortfall av data – se kapitel 11.

Genomför återstående förberedande analyser separat för olika grupper och mättillfällen:

- Identifiera univariata extremvärden (t.ex. med hjälp av box-plot) och helst också multivariata extremvärden (t.ex. med hjälp av hävstångsvärden och DfBeta-värden).
- Undersök univariat normalfördelning (t.ex. genom värden för skevhet och kurtos hos de viktigaste kontinuerliga variablerna) och helst också multivariat normalfördelning (t.ex. med Mardias mått). Ta reda på om spridningen är jämn eller ojämn (med t.ex. Levenes test) och bedöm antagandet om oberoende (gå tillbaka till studiens design och strategi för analys).
- Det finns ingen metod för att hantera de ovan nämnda problemen som inte har nackdelar. Det är därför viktigt att nogsamt väga för- och nackdelar med traditionella respektive nyare lösningar på problem som framkommer i datagranskningen.

Fördjupningslitteratur

Extremvärden:

Barnett, V. & Lewis, T. (1994). *Outliers in statistical data.* New York: Wiley.

Rousseeuw, P. J. & Van Zomeren, B. H. (1990): Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association, 85*, 633–639.

Normalitet:

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2*, 292–307.

Övrigt:

Jaccard, J. & Guilamo-Ramos, V. (2002). Analysis of variance frameworks in clinical child and adolescent psychology: Advanced issues and recommendations. *Journal of Clinical Child Psychology, 31*, 278–294.

Judd, C. M., McClelland, G. H. & Ryan, C. S. (2009). *Data analysis: A model comparison approach*. New York: Routledge.

Maxwell, S. E. & Delaney, H.D. (2004). *Designing experiments and analyzing data: A model comparison perspective*. Mahwah, New Jersey: Erlbaum.

Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. San Diego CA: Academic Press.

Wilcox, R. R. (2005). New methods for comparing group: Strategies for increasing the probability of detecting true differences. *Current Directions in Psychological Science*, *14*, 272-275.

Referenser

Barnett, V. & Lewis, T. (1994). *Outliers in statistical data.* New York: Wiley.

Blaylock, J., Salathe, L. & Green, R. (1980). A note on the Box-Cox transformation under heteroskedasticity. *Western Journal of Agricultural Economics, 45*, 129–135.

Budescu, D. & Appelbaum, M. (1981). Variance stabilizing transformations and the power of the F test. *Journal of Educational and Behavioral Statistics, 6*, 55–74.

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2*, 292–307.

Doksum, K. A. & Wong, C.-W. (1983). Statistical tests based on transformed data. *Journal of the American Statistical Association*, *78*, 411–417.

Everitt, B. S. (2002). *Cambridge dictionary of statistics*. West Nyack, NY: Cambridge University Press.

Jaccard, J. (2006). *Guidelines: Proposal preparation using SEM*. Opublicerat manu-

skript, Florida International University.
Jaccard, J. & Guilamo-Ramos, V. (2002). Analysis of variance frameworks in clinical child and adolescent psychology: Advanced issues and recommendations. *Journal of Clinical Child Psychology, 31*, 278–294.
Kenny, D. & Judd, C. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin*, *99*, 422–431.
Kenny, D. A., Kashy, D. A. & Cook, W. L. (2006). *Dyadic data analysis*. New York: Guilford Press.
Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*, 519–530.
Mardia, K. V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhya*, Series B, *36*, 115–128.
Mardia, K. V. (1980). Tests of univariate and multivariate normality. I P. R. Krishnaiah (Red.), *Handbook of statistics* (Vol. 1, s. 279–320). Amsterdam: North-Holland.
Maxwell, S. E. & Delaney, H.D. (2004). *Designing experiments and analyzing data: A model comparison perspective*. Mahwah, New Jersey: Erlbaum.
Meyers, L. S., Gamst, G. & Guarino, A. J. (2006). *Applied multivariate research: Design and interpretation*. Thousand Oaks, CA: Sage.
Milligan, G. (1987). The use of the arc-sine transformation in the analysis of variance. *Educational and Psychological Measurement, 47*, 563–573.
Parra-Frutos, I. (2009). The behaviour of the modified Levene's test when data are not normally distributed. *Journal Computational Statistics, 24*, 671–693.
Rausch, J. R., Maxell, S. E. & Kelly, K. (2003). Analytic methods for questions pertaining to a randomized pretest, posttest, and follow-up design. *Journal of Clinical Child and Adolescent Psychology, 32*, 467–486.
Raykov, T. & Marcoulides, G. A. (2008). *Introduction to applied multivariate analysis*. New York: Routledge, Taylor and Francis Group.
Ridenour, T. A., Clark, D. B. & Cottler, L. B. (2009). The Illustration-based assessment of liability and exposure to substance use and antisocial behavior for children. *The American Journal of Drug and Alcohol Abuse, 35*, 242–252.
Rousseeuw, P. J. & Van Zomeren, B. H. (1990): Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association, 85*, 633–639.
Tabachnick, B. G. & Fidell, L. S. (2007). *Using multivariate statistics, 5th ed.* Boston: Allyn and Bacon.
Tukey, J. W. (1977). *Exploratory data analysis*. Reading, Mass.: Addison-Wesley.
Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods. *American Psychologist, 53*, 300–314.
Wilcox, R. R. (2002). Understanding the practical advantages of modern ANOVA methods. *Journal of Clinical Child and Adolescent Psychology, 31*, 399–412.
Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. San Diego, CA: Academic Press.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). San Diego, CA: Academic Press.
Wilcox, R. R., Charlin, V. L. & Thompson, K. (1986). New Monte Carlo results on the robustness of the ANOVA F, W, and F* statistics. *Communications in Statistics. Simulation and Computation, 15,* 933–944.
Yuan, K-H. & Hayashi, K. (2010). Fitting data to model: Structural equation modeling diagnosis using two scatter plots. *Psychological Methods, 15*, 335–351.
Zijlstra, W. P., van der Ark, A. & Sijtsma, K. (2007). Outlier detection in test and questionnaire data. *Multivariate Behavioral Research, 42*, 531–555.

Kapitel 13 | Sandra South och Gunnar Bjørnebekk

Variansanalys

Variansanalysen (eng. *Analysis of Variance, ANOVA*) utvecklades på 1920-talet för att användas i analysarbete inom jordbruk. Därefter har variansanalys kommit att användas inom många forskningsområden, både inom grund- och tillämpad forskning. Variansanalys är särskilt väl lämpad för att jämföra grupper av data, till exempel en interventions- och en jämförelsegrupp. Därför används ofta variansanalys vid effektutvärderingar.

Syftet med det här kapitlet är att gå igenom termer och begrepp som ofta förekommer inom variansanalys samt presentera de analystyper som vanligen används vid effektutvärderingar. För den intresserade finns förslag på fördjupningslitteratur i slutet av kapitlet. Ett generellt råd är att vid eventuell osäkerhet kontakta en statistiker vid planeringsstadiet av en effektstudie.

Vad är variansanalys?

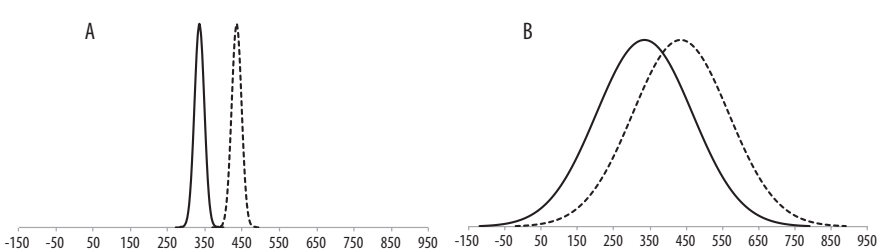
Variationens betydelse

Variansanalys är ett verktyg för att utforska om medelvärden mellan två eller flera grupper av individer skiljer sig signifikant åt. Analysen utgår från två hypoteser. Den första kallas nollhypotesen, och innebär att det inte finns någon skillnad mellan gruppernas medelvärde (H_0). Om nollhypotesen är falsk, och därför kan förkastas, accepte-

ras alternativhypotesen (H_1). En förkastning av nollhypotesen betyder därmed att det finns en statistiskt signifikant skillnad mellan gruppernas medelvärden.

Förutom medelvärde är också spridningen av mätvärden (variationen) ett centralt begrepp inom variansanalys. Anledningen till det kan förklaras med följande exempel: i en randomiserad kontrollerad studie (RCT) angående studenters matteresultat jämförs två skolor (A och B). Hälften av studenterna i varje skola får en intervention och andra hälften utgör jämförelsegrupp och får standardbehandling. Studenternas matteresultat är utfallsvariabeln (eng. *outcome variable*), vilken fungerar som kriterium för att värdera interventionens effekter. Studenternas grupptillhörighet fungerar som den orsaksvariabel (eng. *explanatory variable*) som är avsedd att påverka matteresultat. Eftersom utfallsvariabeln (matteresultat) förväntas bero på orsaksvariabeln (studenternas grupptillhörighet), kallas utfallsvariabeln också för beroendevariabeln (eng. *dependent variable*). Som följd kallas orsaksvariabeln för oberoendevariabeln (eng. *independent variable*).

Nollhypotesen i vårt exempel är att interventionen inte har någon effekt, det vill säga att det inte finns någon statistiskt säkerställd



Figur 13:1. Kurvor som visar medelvärdet och variationen inom och mellan interventions- (streckad linje) och jämförelsegrupp (heldragen linje) grupp i skola A och skola B.

skillnad i studenternas genomsnittliga matteresultat mellan interventions- och jämförelsegruppen. Medelvärdena var samma i både skolorna A och B: 325 i jämförelsegrupperna och 435 i interventionsgrupperna. Topparna på kurvorna i figur 13:1 visar medelvärdena.

Värdenas variation i de två skolorna skiljde sig dock åt trots att medelvärdet är samma. Figur 13:1 visar att variationen mellan individerna i skola B är tio gånger större än variationen mellan individerna i skola A, vilket återspeglas i att kurvans ”svansar” är tio gånger längre. Detta bekräftas av en deskriptiv analys av datamaterialet som visar att standardavvikelsen (ett vanligt spridningsmått som beskriver variationen mellan individer) för båda grupperna i skola A är 13,2 medan standardavvikelsen för båda grupperna i skola B är 132,3. Denna skillnad får stor betydelse vid jämförelsen av medelvärden. En statistisk analys visar att de två grupperna skiljer sig signifikant åt i skola A och att nollhypotesen därmed kan förkastas. Däremot kan analysen inte förkasta nollhypotesen i skola B; i skola B är nämligen skillnaderna mellan individerna mycket större än skillnaden mellan grupperna och det finns därmed inte någon statistiskt säkerställd skillnad mellan grupperna. Hade man enbart tittat på medelvärdena i denna undersökning hade ovan beskrivna skillnader inte upptäckts. Detta är intuitivt förståeligt med tanke på att förändringen på 110 poäng i matteresultat har större betydelse när studenternas mattebetyg utan interventionen (d.v.s. i jämförelsegruppen) förväntas ligga mellan 300 och 350 poäng än när det förväntas ligga mellan 60 och 590 poäng.

En tumregel är att om variationen inom en grupp överlappar den andra gruppens medelvärde (d.v.s. svansen av en kurva överlappar toppen av den andra), så behövs en statistisk analys för att avgöra om det finns en statistiskt signifikant skillnad mellan grupperna.

Hur fungerar variansanalys?

Variansanalys är ett sätt att dela upp variationen i en utfallsvariabel och därigenom testa om minst en grupps medelvärde skiljer sig från de andras. Variationen är uppdelad i två komponenter. Den första

anger variationen mellan grupperna; hur mycket gruppernas medelvärden skiljer sig. Variationen mellan grupperna räknas ut genom att titta på hur mycket varje grupps medelvärde skiljer sig från det totala medelvärdet (d.v.s. totalmedelvärdet för samtliga individer i studien). I exemplet ovan är variationen mellan grupperna identisk i båda skolorna.

Som visades i exemplet ovan, måste skillnaderna mellan medelvärden värderas med hänsyn till variation mellan individer. Denna variation avspeglas i den andra variationskomponenten i en variansanalys: variationen inom grupperna. Den räknas ut genom att titta på hur mycket individernas värden skiljer sig från deras respektive grupps medelvärde. I exemplet ovan är variationen inom grupperna större i skola B än i skola A. Variationen inom grupperna avspeglar hur mycket av variationen som inte kan förklaras av orsaksvariabeln och kallas därför också för residualvarians (eng. *residual variance*) eller felvarians (eng. *error variance*).

Jämförelsen av variationen mellan grupperna med variationen inom grupperna i en variansanalys resulterar i ett *F*-värde som också kallas för en *F*-kvot (eng. *F-value* eller *F-ratio*; uppkallad efter Fisher som utvecklade variansanalys), som formuleras:

$$F = \frac{\textit{variationen mellan grupperna}}{\textit{variationen inom grupperna}}$$

Ett *F*-värde på till exempel fyra visar att variationen mellan grupperna är fyra gånger större än variationen inom grupperna. Ju större variationen inom grupperna är, desto mindre blir *F*-värdet. Statistisk signifikans bedöms med hjälp av en *F*-tabell som visar hur stort *F*-värdet måste vara för att förkasta nollhypotesen. Den kritiska gränsen för när ett *F*-värde är tillräckligt stort för att visa en statistiskt signifikant skillnad mellan grupperna beror på hur många individer som ingår i studien och hur många olika grupper det finns i studien, det vill säga frihetsgraderna (eng. *degrees of freedom*). Signifikansvärdet, kallat *p*-värde (eng. *p-value*, *probability value*) visar sannolikheten att man får ett större *F*-värde än det observerade värdet givet att nollhy-

potesen är sann. Oftast godtas en kritisk gräns av 0,05 för *p*-värdet, vilket innebär att man accepterar att det är fem procents sannolikhet att man får ett *F*-värde som är större än det observerade *F*-värdet och därigenom felaktigt förkastar nollhypotesen (Huberty, 1993).

Med hjälp av ett statistiskt program (t.ex. SPSS eller SAS) är det lätt att genomföra en variansanalys. Resultaten är också enkla att tolka eftersom de är upplagda i en tabell som redovisar mått på variation, *F*-värdet och signifikansnivån. Problemet med variansanalys är att välja rätt typ av variansanalys.

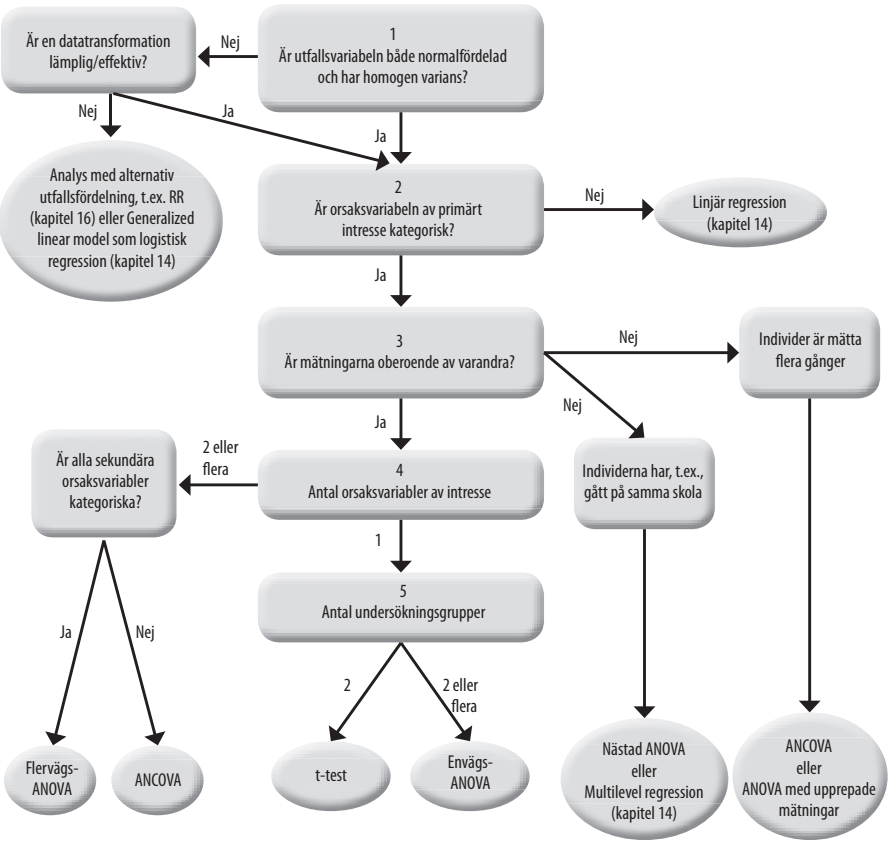
Förutsättningar för variansanalys

Det är viktigt att noggrant tänka igenom vilken typ av analys som passar den fråga man vill undersöka innan datainsamlingens början. Datamaterial som passar enklare typer av analyser baserade på noggrann experimentell design (t.ex. flervägs-ANOVA eller ANCOVA) är önskvärt av två anledningar. För det första är det lättare att genomföra och tolka analysen. För det andra, och viktigast, är experimentell kontroll *alltid* mer effektivt än statistisk kontroll och i vissa fall finns det inte något sätt att kontrollera för oönskad variation på grund av ovidkommande (eng. *confounding*) faktorer.

Det finns flera olika typer av variansanalyser och som stöd för att välja analystyp kan man använda de faktorer som åskådliggörs i beslutsdiagrammet i figur 13:2. Diagrammet omfattar inte samtliga möjliga frågor som rör val av variansanalys eller samtliga möjliga variansanalystyper, utan fokuserar på några som ofta används i effektutvärderingar. Börja med fråga 1 och fortsätt svara på frågorna (fyrkanter) tills du når en analystyp (oval).

De första tre frågorna i beslutsdiagrammet rör fyra av de antaganden som måste uppfyllas för alla typer av variansanalys (jfr kapitel 12):

1. att utfallsvariabeln är normalfördelad
2. att utfallsvariabeln har homogen varians
3. att orsaksvariabeln av primärt intresse är kategorisk
4. att mätningarna är oberoende av varandra.

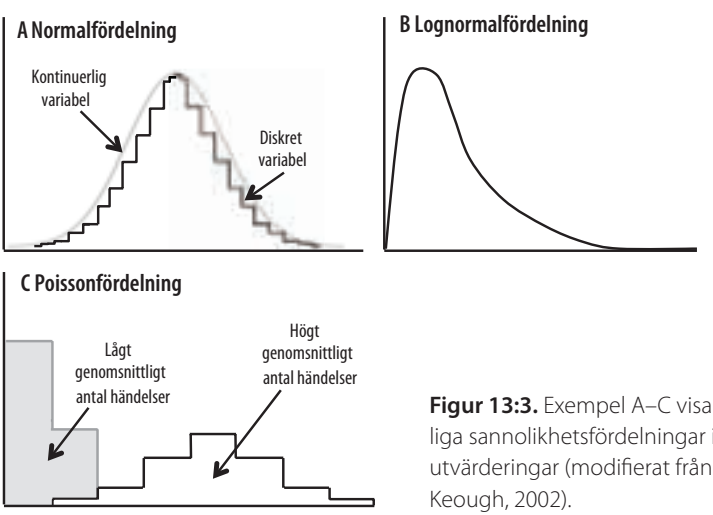


Figur 13:2. Beslutsdiagram för val av analystyp.

Dessa antaganden kommer att beröras ytterligare i de följande styckena och ligger till grund för valet av variansanalys. Statistiker och forskare är dock inte överens om hur robust variansanalysen är för avvikelser från dessa antaganden. Däremot verkar de vara överens om att effekten av avvikelser från dessa antaganden är små i jämförelse med metodfel i design och genomförande (t.ex. bristande randomisering). En skattning av den relativa betydelsen av antagandena sammanfattas av Brown, Costigan och Kendziora (2008).

Normalfördelad utfallsvariabel

Den första delen av fråga 1 i beslutsdiagrammet handlar om det första antagandet: att utfallsvariabeln är normalfördelad. Ett kännetecken på normalfördelning är symmetrisk variation kring medelvärdet, vilket ger en karakteristisk klockform. I figur 13:3 presenteras tre exempel på vanliga fördelningar inom effektutvärderingar. Exempel A i figuren visar den önskvärda klockformade normalfördelningen. Normalfördelningen beskriver en kontinuerlig variabel som kan anta alla tänkbara värden, till exempel längd mätt i cm med tre decimaler. Det är dock extremt ovanligt att uppnå denna precisionsnivå och man använder istället en mer hanterbar skala som bara kan anta vissa värden/skalsteg (t.ex. längd i cm avrundat till närmaste heltal), en så kallad diskret variabel. Skillnaderna mellan dessa skalsteg har samma innebörd oavsett var på skalan man befinner sig (s.k. ekvidistanta skalsteg). Exempel A visar hur sådana diskreta variabler kan likna en normalfördelning när mätskalan kan bestå av många tänkbara ekvidistanta skalsteg. Detta betyder att diskreta variabler med många tänkbara ekvidistanta skalsteg kan betraktas som kontinuerliga och kan därför analyseras med en variansanalys.



Figur 13:3. Exempel A–C visar några vanliga sannolikhetsfördelningar inom effektutvärderingar (modifierat från (Quinn & Keough, 2002).

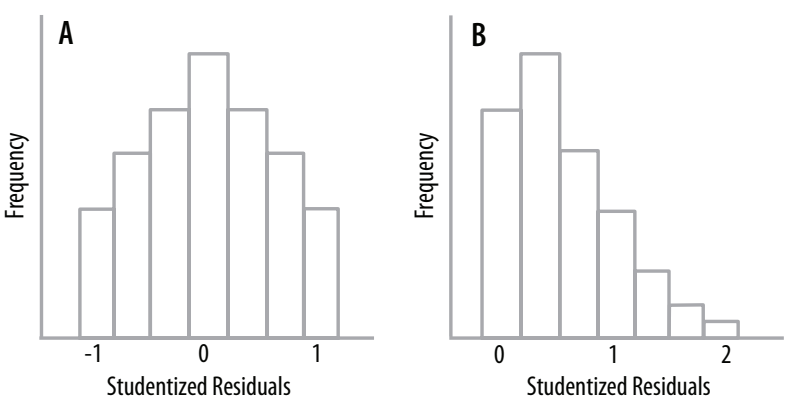
När variansen kring medelvärden av en kontinuerlig variabel är osymmetrisk kan det resultera i en *lognormalfördelning*, som exempel B i figur 13:3. En sådan fördelning kan exempelvis uppstå om 75 procent av individerna får ett matteresultat under 40 poäng, 15 procent får ett resultat mellan 40 och 60 poäng och kvarstående 10 procent är utspridda mellan 60 och 200 poäng. Denna fördelning är därför skev (eng. *skewed*) åt höger, vilket ses som en långsmal ”svans” åt höger. Om utfallsvariabeln liknar en lognormalfördelning kan den transformeras så att den blir mer lik en normalfördelning. Transformationsprocedurer är omdiskuterade och bör endast genomföras efter noggrant övervägande (jfr kapitel 12). Om det inte går att transformera utfallsvariabeln kan datamaterialet analyseras med, till exempel, en *Generalized Linear Model* som möjliggör jämförelser av grupper med icke-normala utfallsfördelningar (Dobson, 2001; McCullagh & Nelder, 1989).

I effektutvärderingar är det vanligt att räkna antalet gånger en händelse uppstår för varje individ (t.ex. hur många gånger man blir registrerad i lagföringsregistret under en uppföljningsperiod). Antalet registreringar blir då en diskret variabel med ekvidistanta skalsteg. Ett annat kännetecken för en sådan variabel är att den inte kan anta negativa tal eftersom en händelse inte kan inträffa mindre än noll gånger. En poissonfördelning beskriver bäst fördelningen av sådana variabler. Den vita kurvan i exempel C åskådliggör hur en poissonfördelning börjar likna en normalfördelning när utfallsvariabeln har ett högt genomsnittligt antal händelser (d.v.s. händelsen är vanlig) och kan då analyseras med en variansanalys. Om det genomsnittliga antalet händelser däremot är lågt (d.v.s. händelsen är sällsynt) och fördelningen därmed snedställd (som i den grå kurvan) kan datamaterialet analyseras med, till exempel, en *Generalized Linear Model* (Dobson, 2001; McCullagh & Nelder, 1989).

Utfallsvariabeln är dikotom när man mäter om en händelse inträffat eller inte, exempelvis återfall i brottslighet (d.v.s. minst en registrering i lagföringsregistret). Då handlar det om en ordinal eller nominal variabel. En sådan fördelning kallas en binomialfördel-

ning och bör inte analyseras med variansanalys. Det finns många olika typer av analyser som är lämpliga för dikotoma variabler, till exempel relativ risk (eng. *risk ratio*, RR) (kapitel 16), logistik regression (kapitel 14) eller någon annan typ av *Generalized Linear Model* anpassad för binomiala utfallsfördelningar (Dobson, 2001; McCullagh & Nelder, 1989).

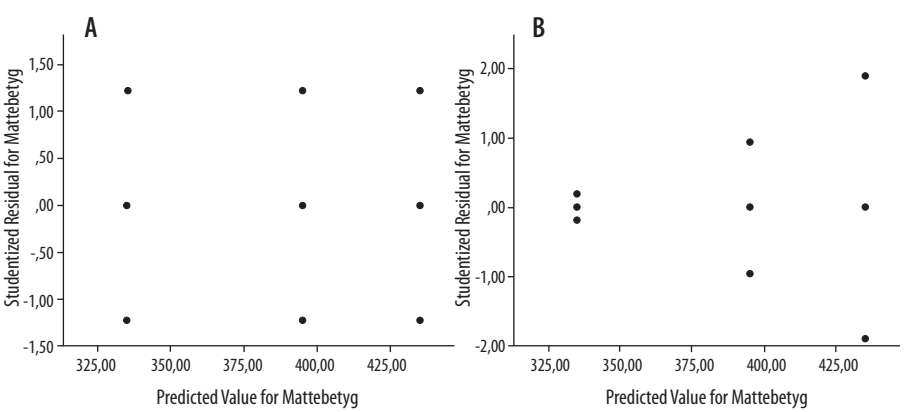
Avvikelser från en normalfördelning kan testas med hjälp av diagnostiska analyser (se kapitel 12). Det första steget bör dock vara att skapa ett diagram av utfallsvariabeln för att lära känna datamaterialet. Utfallsvariabelns fördelning kan snabbt och enkelt kontrolleras i de flesta statistiska program genom att klicka på ”spara residualer” (eng. *save residuals*). Då sparas ett värde för varje individ vilket representerar hur långt individen ligger från sin grupps medelvärde (därigenom kopplingen mellan termerna residualvarians och variation inom grupperna). Med hjälp av dessa residualvärden är det möjligt att skapa ett histogram för varje grupp och därmed se hur spridningen av variationen inom varje grupp fördelar sig kring medelvärdet (ett residualvärde på 0 betyder att individen ligger på medelvärdet). Figur 13:4 exemplifierar hur residualerna kan se ut för en normalfördelad respektive en lognormalfördelad utfallsvariabel.



Figur 13:4. Exempel på histogram av residualer från en variansanalys för a) en normalfördelad utfallsvariabel och b) en lognormalfördelad utfallsvariabel.

Homogen varians

Den andra delen i fråga 1 i beslutsdiagrammet berör antagandet om utfallsvariabeln har homogen varians. Antagandet innebär att variationen inom grupperna bör vara lika stor för alla grupper. I annat fall är variansen heterogen, något som exempelvis kan bli en följd av både skillnader i antalet individer i de undersökta grupperna och om utfallsvariabeln inte är normalfördelad (Coombs, Algina & Oltman, 1996). Det finns diagnostiska analyser för att undersöka om varianserna är homogena (se kapitel 12), men rekommendationen är att först testa detta antagande genom att visuellt utforska variationens spridning. Detta kan göras genom att ”spara residualer” i de flesta program. För att undersöka om variationen mellan grupperna är homogen måste man också ”spara predicerade värden” (eng. *save predicted values*). Detta skapar två kolumner till i datafilen; den ena med varje individs residualvärde och den andra med deras respektive grupps medelvärde. Ett spridningsdiagram (eng. *scatterplot*) som ser ut som det i figur 13:5 kan då skapas.



Figur 13:5. Exempel på ett spridningsdiagram av predicerade värden i förhållande till residualer för homogen varians (A) och heterogen varians (B, notera trattformen).

Kategorisk orsaksvariabel av primärt intresse

Det tredje antagandet som måste uppfyllas inför variansanalys är att orsaksvariabeln av primärt intresse är uppdelad i kategoriska grupper (t.ex. interventions- och kontrollgruppen) (figur 13:2, fråga 2). Om orsaksvariabeln av primärt intresse däremot är kontinuerlig kan variansanalys inte användas och en regressionsanalys är att föredra (se kapitel 14).

Mätningarna är oberoende av varandra

Det fjärde antagandet som måste uppfyllas för variansanalys är att mätningarna är oberoende av varandra (figur 13:2, fråga 3). När det finns ett samband mellan de individuella utfallsvärdena är mätningarna beroende av varandra. Något som kan uppstå när till exempel samma individ är mätt upprepade gånger. Utfallet vid eftermätningen är därigenom beroende av utfallet vid förmätningen. När randomisering görs på gruppnivå i stället för individnivå kan också ett beroende mellan mätningarna uppstå; i detta fall på grund av att en individs utfall kan påverka utfallet för en annan individ i gruppen.

Om inte hänsyn tas till eventuellt beroende mellan mätningar är risken att jämförelserna mellan grupperna blir felaktiga och att en sann nollhypotes därigenom förkastas (typ I-fel). I vissa fall går det att kontrollera för avvikelser från detta antagande genom experimentell design och val av rätt typ av variansanalys (se Analys av mer verklighetsnära datamaterial).

Enklare analystyper

I följande avsnitt presenteras två enklare analystyper som endast är lämpliga med perfekt randomisering av deltagarna och när den främsta källan till systematiska skillnader mellan individernas utfall beror på vilken grupp individerna ingår i. Sådana förutsättningar finns sällan i realiteten eftersom de utfallsmått som vanligen används nästan alltid även är påverkade av andra variabler (t.ex. bakgrundsfaktorer). Däremot är det viktigt att förstå dessa enkla-

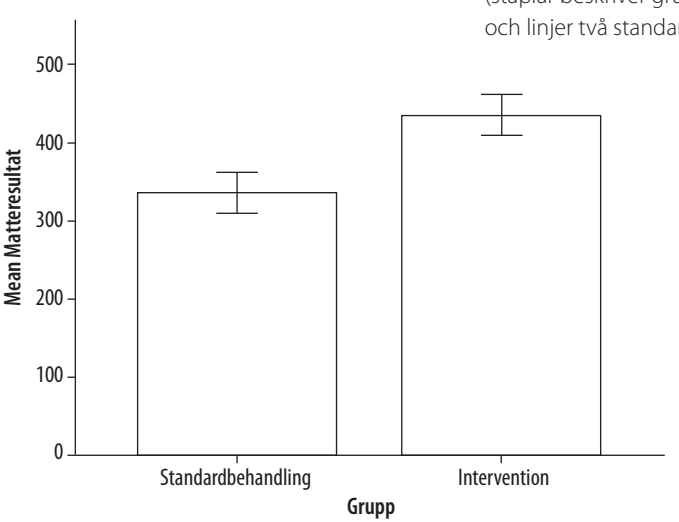
re analystyper eftersom de tydliggör de grundläggande principerna bakom variansanalys.

T-test

När det bara finns en orsaksvariabel (figur 13:2, fråga 4) och två grupper (figur 13:2, fråga 5) så är ett oberoende tvågruppers *t*-test (eng. *independent two-sample t-test*) lämpligt.

Tekniskt sett är ett *t*-test inte en variansanalys eftersom det inte är baserat på ett *F*-värde utan på ett *t*-värde. Däremot liknar *t*-värdet och *F*-värdet varandra eftersom båda jämför skillnaden mellan medelvärden samtidigt som de tar hänsyn till variationen mellan individer. Ett *t*-test av skola A:s datamaterial visar att interventionen hade en statistiskt signifikant effekt på Matteresultat eftersom signifikansvärdet (*p*-värdet) är mindre än 0,05 (figur 13:6)[1]. I figuren redovisas även medelvärde och standardavvikelse i skola A i form av

	t	d.f.	Signifikans
Matteresultat	-9,26	4	0,001



Figur 13:6. *T*-test av skillnaden i Matteresultat mellan grupperna (staplar beskriver gruppmedelvärdet och linjer två standardavvikelser).

1 Se bilaga 1A för datamaterialet till exemplet.

ett stapeldiagram. Skillnaden mellan staplarna visar att personerna i jämförelsegruppen som fått standardbehandlingen har ett lägre genomsnittligt matteresultat än de i interventionsgruppen. Figur 13:6 visar att variationen inom en grupp (vertikal linje) inte överlappar den andra gruppens medelvärde (stapelns höjd). Det betyder att medelvärdena inte ligger inom två standardavvikelser från varandra.

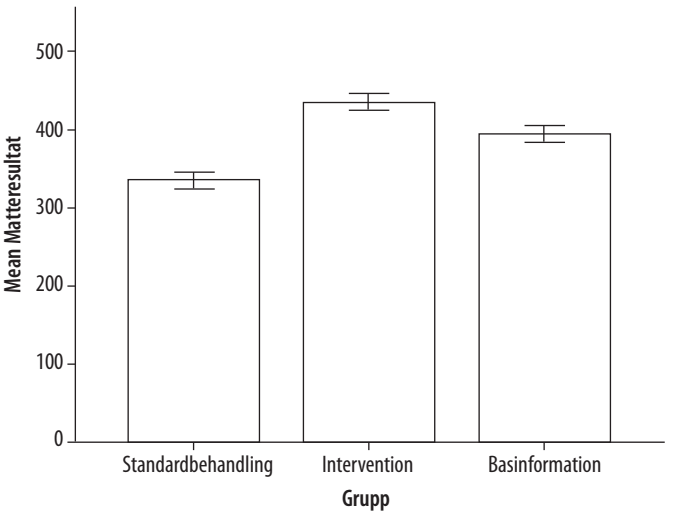
Envägs-ANOVA

Envägs-ANOVA (eng. *one-way* ANOVA) används när det finns en orsaksvariabel (figur 13:2, fråga 4) och fler än två grupper (figur 13:2, fråga 5). Som tidigare beskrevs, avgörs en statistisk skillnad mellan grupperna i en ANOVA med hjälp av *F*-värdet, som är variationen mellan grupperna dividerat med variationen inom grupperna.

Figur 13:7 visar resultatet av en envägs-ANOVA med tre grupper[2]. Mean Square-värdet för ”Behandling” (eng. *treatment*) representerar variationen mellan grupperna. Mean Square-värdet för ”Error” (residualvarians) representerar variationen inom grupperna (för beräkning av dessa värden manuellt, se Borg & Westerlund, 2010; Hassmén & Koivula, 1996). Genom att jämföra den andra och fjärde kolumnen i figur 13:7 åskådliggörs att *F*-värdet är lika med Mean Square-värdet för ”Behandling” dividerat med Mean Square-värdet för ”Error” (d.v.s. $F = 7600/25$). Signifikansnivån med 2 och 6 frihetsgrader är mindre än 0,001 och visar därför att åtminstone en grupp skiljer sig signifikant från de andra. Granskning av figur 13:7 visar att inget av de tre medelvärdena ligger inom två standardavvikelser från en annan grupp. Det betyder att alla tre grupper skiljer sig signifikant åt, vilket innebär att det inte finns något behov av vidare analys. Individerna som fick interventionen hade i genomsnitt det högsta matteresultatet, följt av individerna i basinformationsgruppen. Lägst medelvärde hade gruppen som fick standardbehandling.

2 Se bilaga 1B för datamaterialet till exemplet.

	Mean Square	d.f.	F	Sig.
Behandling	7 600	2	304,00	0,000
Error	25	6		



Figur 13:7. Envägs-ANOVA av skillnaden i Matteresultat mellan grupperna (staplar beskriver gruppmedelvärdet och linjer två standardavvikelser).

Om ett eller flera medelvärden hade legat inom två standardavvikelser från varandra (d.v.s. att linjerna överlappat staplarnas höjd) hade det funnits behov av fortsatt analys av datamaterialet med hjälp av *post hoc*-analyser (post hoc = efterhand). Detta innebär en jämförelse av varje par av grupper (eng. *multiple comparisons*) för att identifiera var den signifikanta skillnaden finns. Det är dock en omdiskuterad aspekt av variansanalys (för mer information se Hancock & Klockars, 1996; Hochberg & Tamhane, 1987; Kirk, 1995) och behandlas inte vidare i detta kapitel.

Variansanalys av mer verklighetsnära datamaterial

Variation i utfallet orsakas oftast av flera variabler än interventionen (t.ex. bakgrundsfaktorer). Sådana källor av variation minskar sannolikheten att upptäcka en sann skillnad mellan grupperna, det

vill säga en effekt av interventionen. I denna sektion presenteras hur ovidkommande variation kan kontrolleras för genom experimentell design och variansanalys.

Genom att reducera variationen inom grupperna ökas sannolikheten att upptäcka en sann skillnad mellan grupperna (d.v.s. öka den statistiska styrkan, eng. *power*, eller sensitiviteten; kapitel 4). Variation inom grupperna speglar experimentella fel (t.ex. variation på grund av olika intervjuare) och individuella skillnader (t.ex. variation mellan individer som beror på bakgrundsfaktorer). Ovidkommande variation inom grupperna kan minskas genom att inkludera sekundära orsaksvariabler i den experimentella designen och analysen. Val av sekundära orsaksvariabler bör baseras på teori och tidigare analyser. Det finns till exempel ingen anledning att inkludera hårfärg som en sekundär variabel i en studie om inlärningsmetoder för matte, om hårfärgen inte är förväntad att påverka matteresultatet. Sekundära orsaksvariabler som bidrar till ovidkommande variation kan vara både kategoriska och kontinuerliga.

Under planeringsstadiet av en RCT måste man överväga om systematiska skillnader mellan grupperna kan uppstå trots randomisering och noggrant val av sekundära variabler. Variansanalys är inte lämplig när dessa källor till systematiska skillnader mellan grupperna inte har förutsetts och därför inte heller har inkluderats som en del av den experimentella designen. Om det misstänks att det finns stora skillnader mellan grupperna kan regressionsanalys vara ett alternativ (se kapitel 14) eller en *Propensity Score Analysis* (Guo & Fraser, 2010). På så sätt ändras frågeställningen till att utforska sambanden mellan variablerna i stället för att testa skillnaderna mellan grupperna (som i variansanalys). Det är viktigt att betona att det i vissa fall inte finns något statistiskt verktyg som kan kontrollera för systematiska fel och en bra experimentell design är därför centralt.

Flervägs-ANOVA (faktoriell)

Flervägs-ANOVA (eng. *multi-factor ANOVA*) används när det finns fler än en kategorisk orsaksvariabel (d.v.s. faktor) (figur 13:2, fråga

		Faktor B- Föräldrars utbildning		
		Grundskola	Gymnasium	Högskola
Faktor A- Behadnling	Standardbehandling	n = 2	n = 2	n = 2
	Intervention	n = 2	n = 2	n = 2
	Basinformation	n = 2	n = 2	n = 2

Figur 13:8. Ett exempel på hur en faktoriell design är upplagt.

4). Inkluderandet av sekundära kategoriska orsaksvariabler av intresse minskar inte endast ovidkommande variation i utfallet utan möjliggör också en undersökning av om interventionen har samma effekt på olika klientgrupper. Till exempel om interventionen har olika effekt för flickor och pojkar.

Den viktigaste förutsättningen för flervägs-ANOVA är att den experimentella designen är faktoriell (eng. *factorial*). I figur 13:8 har en kategorisk orsaksvariabel, nämligen Föräldrars utbildning, lagts till i det tidigare exemplet, för att se om interventionen ger samma effekt oavsett föräldrars utbildningsnivå. Notera att det finns individer i varje kombination av de två faktorerna av intresse. Faktoriella designer kallas därför också för *ortagonala* eller *korsade* designer (eng. *orthogonal designs* eller *crossed designs*).

Datamaterialet analyseras med hjälp av flervägs-ANOVA (tabell A, figur 13:9)[3], som i detta fall även kallas för tvåvägs-ANOVA eftersom det finns två kategoriska orsaksvariabler. Det finns en statistiskt signifikant effekt av både Behandling och Föräldrars utbildning ($p < 0{,}05$). Tabell A redovisar också ett F-värde för en typ av effekt som inte har setts hittills: ”Behandling x Föräldrars utbildning”. Denna interaktionseffekt visar hur de kategoriska orsaksvariablerna interagerar med varandra och belyser om interventionen har samma effekt oavsett föräldrarnas utbildningsnivå. Den signifikanta interaktionen i figur 13:9 visar att effekten av interventionen

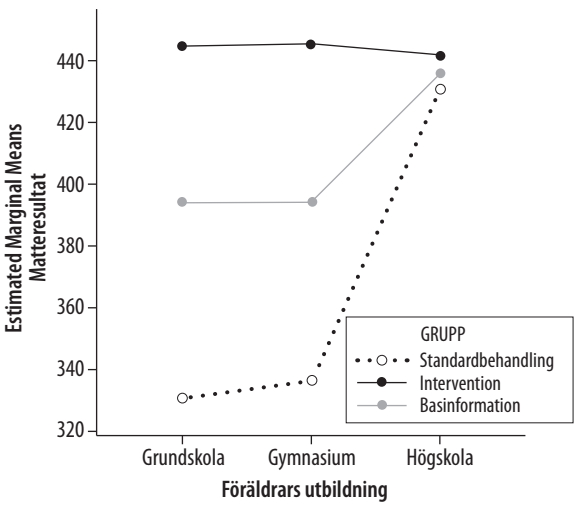
3 Se bilaga 1C för datamaterialet till exemplet.

A

	Mean Square	d.f.	F	Sig.
Behandling	9184	2	344,40	0,000
Föräldrars utbildning	4118	2	154,44	0,000
Behandling x Föräldrars utbildning	1703	4	63,87	0,000
Error	27	9		

B

	Mean Square	d.f.	F	Sig.
Behandling	9184	2	9,01	0,003
Error	1019	15		



Figur 13:9. Flervägs faktoriell ANOVA av effekten av Behandling och Föräldrars utbildning på Matteresultat (A) samt ANOVA av samma datamaterial utan inkluderandet av Föräldrars utbildning (B).

delvis beror på föräldrars utbildningsnivå. En figur av de beräknade medelvärdena för grupperna efter kontroll för orsaksvariablerna i analysen (som kallas för *Estimated Marginal Means* i SPSS) visar att interventionen har påtagligt större effekter när föräldrarna har en grundskole- eller gymnasieutbildning än när de har en högskoleutbildning (figur 13:9).

Om datamaterialet analyseras i en envägs-ANOVA utan orsaksvariabeln Föräldrars utbildning blir variationen *mellan* grupperna

(d.v.s. Mean Square för Behandling) densamma (tabell B, figur 13:9). Det som skiljer är variationen *inom* grupperna (Mean Square-värdet för Error). Detta resulterar i ett lägre *F*-värde när effekten av Föräldrars utbildning inte kontrolleras. Det visar dels att effekten av interventionen beror på föräldrars utbildningsnivå, dels att den ovidkommande variationen har minskat inom grupperna och att analysen kommit närmare försöket att isolera behandlingseffekten.

Ibland är det omöjligt att få lika många individer i varje kombination av de två orsaksvariablerna (figur 13:8) och datamaterialet blir obalanserat (eng. *unbalanced*). I sådana fall rekommenderas ofta flervägs-ANOVA med en specifik beräkningstyp av variationsmåtten (nämligen Type III Sums of Squares; se Milliken & Johnson, 1984; för avvikande uppfattning se Langsrud, 2003). Om vissa celler helt saknar individer är designen inte längre faktoriell eller helt korsad och detta representerar då ett fall av obalanserad data som har mer allvarliga konsekvenser eftersom interaktionseffekten inte kan kalkyleras. I sådana fall rekommenderas regressionsanalys.

ANCOVA med oberoende mätningar

Kovariansanalys (eng. *analysis of covariance, ANCOVA*) är lämplig när det finns fler än en orsaksvariabel av intresse och minst en är kontinuerlig (figur 13:2, fråga 4). Sådana sekundära kontinuerliga orsaksvariabler kallas för kovariater (eng. *covariate*) och inkluderas för att minska ovidkommande variation och för att utforska om andra variabler än interventionen påverkar utfallet. Generellt sett kan man säga att i en ANCOVA justeras först utfallsvariabeln utifrån effekten av kovariatet. Sedan genomförs en ANOVA på den justerade utfallsvariabeln för att testa om grupperna skiljer sig åt efter kontroll för effekten av kovariatet (Quinn & Keough, 2002, s. 342–347).

Utöver de förutsättningar som gäller för variansanalys som diskuterats hittills finns ytterligare tre viktiga antaganden som behöver uppfyllas för en ANCOVA:

1. Sambandet mellan kovariatet och utfallsvariabeln är linjärt (eng. *linearity*).

2. Spridningen av kovariatet är lika för alla grupper (eng. *covariate values similar across groups*).
3. Sambandet mellan kovariatet och utfallsvariabeln är lika för alla grupperna (eng. *homogeneity [equality] of slopes*).

Det första antagandet är att det bör finnas ett linjärt samband mellan kovariatet och utfallsvariabeln (Brown m.fl., 2008; Quinn & Keough, 2002). Detta betyder att en ökning i kovariatet alltid samvarierar med en ökning (eller minskning) av utfallsvariabeln. Med andra ord betyder detta att om man ritade en linje som beskrev sambandet mellan kovariatet och utfallsvariabeln (d.v.s. en regressionslinje), skulle den vara rak.

Det andra antagandet är att värdena för kovariatet är lika för alla grupper; att kovariatets spridning är lika för alla grupperna (Miller & Chapman, 2001; Quinn & Keough, 2002, s. 342–349). Skillnader mellan grupperna i kovariatet uppstår oftast när randomisering av individer till grupperna inte har lyckats eller i icke-randomiserade designer (t.ex. att det finns en skillnad i genomsnittlig ålder mellan interventions- och jämförelsegruppen). Avvikelser från detta antagande resulterar i att sambandet (regressionslinjen) mellan utfallsvariabeln och kovariatet extrapoleras utöver de observerade värdena i vissa grupper. Utfallsvärdet beräknas då för värden på kovariatet där några mätningar inte har gjorts, vilket innebär att viktig variation mellan grupperna försvinner och att utfallsvariabeln därigenom förvrängs (Jamieson, 1999; Miller & Chapman, 2001; Quinn & Keough, 2002). Därför betraktas ANCOVA som olämplig när det finns skillnader mellan grupperna i kovariatet (Jamieson, 1999; Miller & Chapman, 2001; Quinn & Keough, 2002).

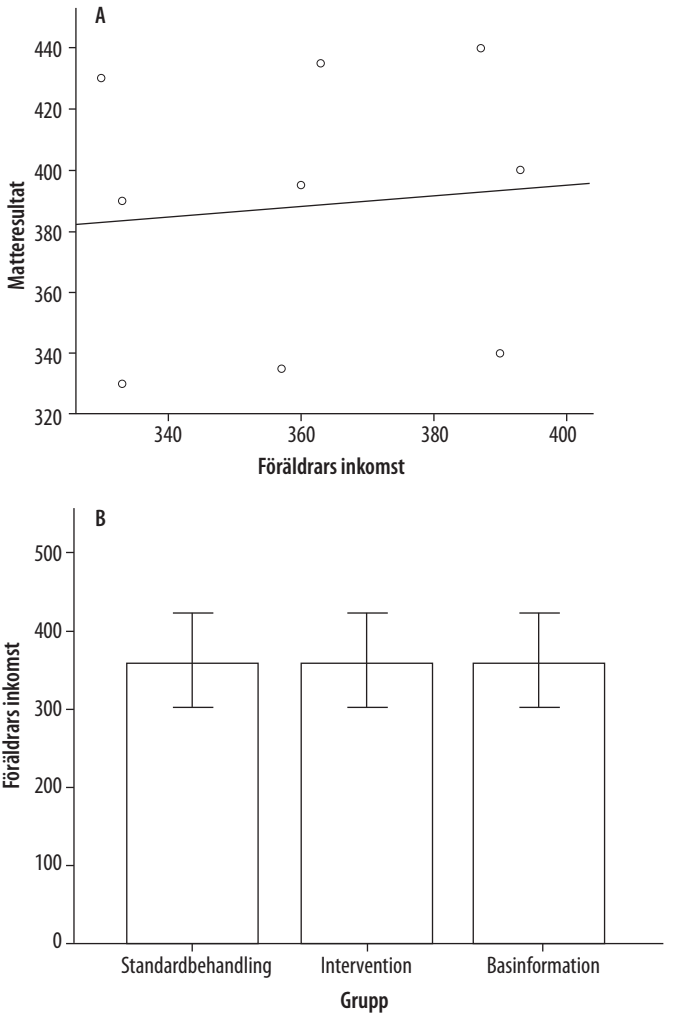
Somliga rekommenderar att ANCOVA endast bör användas när skillnaderna mellan grupperna i kovariatet beror på slumpen (som kan uppstå i en RCT med ett litet antal individer) (Maxwell & Delaney, 1990; Senn, 2006; för avvikande uppfattning se Miller & Chapman, 2001; Quinn & Keough, 2002). I dessa fall är det viktigt att kunna styrka att det inte finns något samband mellan en individs

kovariatvärde och vilken grupp de tilldelades (t.ex. att forskaren inte påverkats av individernas bakgrund när hon fattade ett beslut om grupptilldelning eller att individen inte fick välja grupp själv). Med andra ord betyder detta att man måste kunna argumentera för att det är lämpligt att extrapolera sambandet mellan utfallsvariabeln och kovariatet och att variationen mellan grupper som försvinner efter justering för kovariatet därför inte påverkar behandlingseffekten på utfallsvariabeln. Om det misstänks att det finns systematiska skillnader mellan grupperna på grund av ovidkommande faktorer går det inte att isolera interventionseffekten eller dra slutsatser om kausala samband. Studiens frågeställning, vilken avgör val av analys, blir i stället att utforska de olika sambanden mellan variablerna med hjälp av, till exempel, regressionsanalys (kapitel 14) eller *Propensity Score Analysis* (Guo & Fraser, 2010). Oavsett vilken analystyp som väljs är en stringent tillämpad RCT det bästa sättet att kontrollera för systematiska skillnader som kan uppstå mellan grupperna och bör därför eftersträvas.

Det tredje antagandet som måste uppfyllas inför ANCOVA är att sambandet mellan kovariatet och utfallsvariabeln (d.v.s. lutningen av regressionslinjen) är lika för alla grupper (*Homogeneity [equality] of slopes*; Brown m.fl., 2008; Quinn & Keough, 2002). Det betyder att regressionslinjen som beskriver sambandet mellan kovariatet och utfallsvariabeln måste ha samma lutning i alla grupper. Detta antagande kan testas genom att inkludera interaktionseffekten av kovariatet och Behandling i en ANCOVA.

I figur 13:10 presenteras ett exempel där datamaterialet uppfyller både det första och andra antagandet; ett spridningsdiagram visade att sambandet mellan kovariatet Föräldrars inkomst och Matteresultat är linjärt (figur 13:10 A) och en envägs-ANOVA visade ingen skillnad mellan grupperna i kovariatet Föräldrars inkomst (figur 13:10 B; $F_{2,6} = 0{,}01$, $p = 0{,}995$; nedskänkta siffror visar frihetsgraderna)[4].

4 Se bilaga 1B för datamaterialet till exemplet.

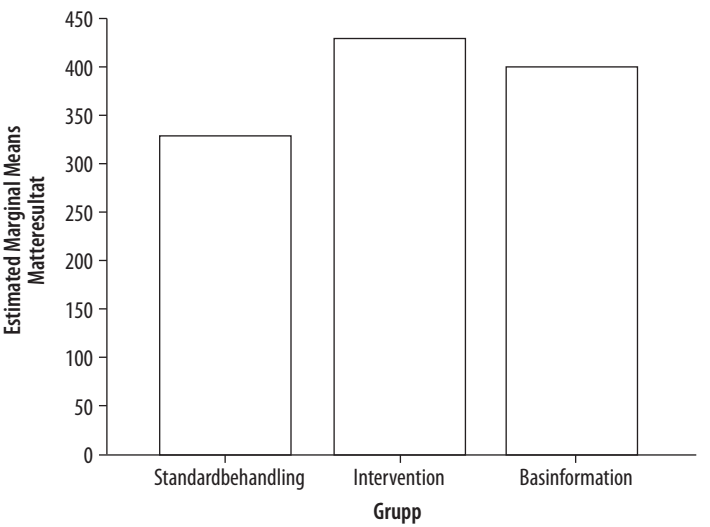


Figur 13:10. Spridningsdiagram med regressionslinje för att visa det linjära sambandet mellan kovariatet Föräldrars inkomst och Matteresultat (A). Medelvärden visar hur Föräldrars inkomst inte skiljer sig mellan grupperna (B).

Det tredje antagandet undersöks med hjälp av en ANCOVA med Matteresultat som utfallsvariabel, Behandling som den primära orsaksvariabeln och Föräldrars inkomst som kovariat samt interaktionen mellan Behandling och Föräldrars inkomst. Resultaten (figur 13:11) visar att det tredje antagandet är uppfyllt eftersom det

	Mean Square	d.f.	F	Sig.
Föräldrars inkomst	149	1	451,15	0,000
Behandling	32	2	97,42	0,002
Behandling x Föräldrars inkomst	0,04	2	0,11	0,9
Error	0,33	3		

Figur 13:11. ANCOVA av effekten av Behandling och kovariatet Föräldrars inkomst med interaktionen mellan Behandling och Föräldrars inkomst som en test för antagandet "Homogeneity of slopes".



inte finns en signifikant interaktion mellan Behandling och Föräldrars inkomst. Variationen kring dessa medelvärden är så pass liten (standardfel = 0,33) att den inte syns. Notera att man måste utelämna den icke-signifikanta interaktionstermen från den slutgiltiga modellen, under förutsättningen att det andra antagandet för ANCOVA är uppfyllt.

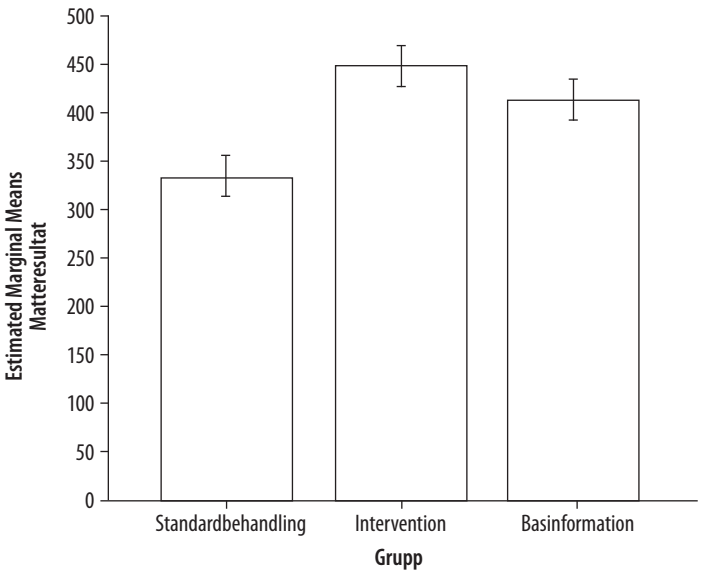
Den slutgiltiga ANCOVA:n är uppbyggd på samma sätt som den förra, fast utan interaktionseffekten. Den visar en statistiskt signifikant effekt av Behandling på Matteresultat ($F_{2,5} = 35730, p < 0{,}001$) och ett statistiskt signifikant samband mellan Föräldrars inkomst och Matteresultat ($F_{1,5} = 701, p < 0{,}001$). Variationen inom grupper-

A

	Mean Square	d.f.	F	Sig.
Matteresultat förmätning	73 816	1	69,10	0,000
Behandling	9 442	2	8,84	0,023
Error	1 068	5		

B

	Mean Square	d.f.	F	Sig.
Behandling	9 448	2	0,72	0,526
Error	13 193	6		



Figur 13:12. ANCOVA av effekten av Behandling på Matteresultat vid eftermätning med Matteresultat vid förmätning som kovariat (A) samt ANOVA av samma datamaterial utan inkluderande av kovariatet (B).

na har minskat eftersom Mean Square-värdet för Error med Föräldrars inkomst som kovariat är mindre (MS_{error} = 0,213) än motsvarande värde i envägs-ANOVA:n (figur 13:7, MS_{error} är 25). De beräknade medelvärdena efter justering för kovariatet (s.k. *Estimated Marginal Means*) ger information om interventionseffekten (figur 13:11).

ANCOVA med upprepade (beroende) mätningar

När en och samma individ undersöks vid flera tillfällen (t.ex. före och efter interventionen) uppstår ett beroende mellan mätningar-

na (figur 13:2, fråga 3). Utan kontroll för ett beroende mellan mätningarna är risken att jämförelserna mellan grupperna blir felaktiga och att en sann nollhypotes därigenom förkastas (typ I-fel). Detta på grund av överskattning av antalet individer i studien, samtidigt som sambandet mellan individerna *inom* grupperna ignoreras. ANCOVA är ett sätt att kontrollera för detta beroende genom att kontrollera för skillnader mellan individer som fanns redan innan studiens start. Genom att ta hänsyn till skillnader mellan individer minskas även ovidkommande variation. Frågeställningen blir då om det finns en interventionseffekt på eftermätningen efter justering för förmätningsvärdet. Förmätningen inkluderas som en kontinuerlig sekundär orsaksvariabel (d.v.s. ett kovariat) i en ANCOVA; på precis samma sätt som Föräldrars inkomst inkluderades som ett kovariat i figur 13:11.

Eftersom den underliggande analysen i ANCOVA med upprepade (beroende) mätningar är densamma som i en ANCOVA med oberoende mätningar, måste samma antaganden vara uppfyllda. I det följande exemplet analyseras ett datamaterial på samma sätt som exemplet i figur 13:11, men nu används Matteresultat vid förmätning som kovariat.[5] Exemplet tydliggör att det är samma tillvägagångssätt som gäller oavsett om kovariatet är en förmätning eller en annan sekundär kontinuerlig orsaksvariabel.

I ett spridningsdiagram framgår att sambandet mellan Matteresultat vid förmätning och Matteresultat vid eftermätning är linjärt; första antagandet inför ANCOVA är därmed uppfyllt. Det andra ANCOVA-antagandet är också uppfyllt; värdena för kovariatet skiljer sig inte mellan grupperna (envägs-ANOVA: $F_{2,6} = 0{,}00$, $p = 1{,}0$). En ANCOVA som inkluderade interaktionen mellan Matteresultat vid förmätning och Behandling visade att det inte fanns en statistiskt signifikant interaktionseffekt och att datamaterialet därför uppfyllde det tredje ANCOVA-antagandet ($F_{2,3} = 0{,}02$, $p = 0{,}984$).

Figur 13:12 (tabell A) visar ett statistiskt signifikant samband

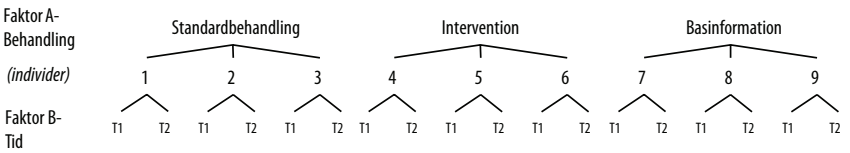
5 Se bilaga 1D för datamaterialet till exemplet.

mellan Matteresultat vid förmätningen och Matteresultat vid eftermätningen, samt en statistiskt signifikant effekt av Behandling. För jämförelsens skull analyseras samma datamaterial med hjälp av en envägs-ANOVA utan Matteresultat vid förmätning som kovariat (tabell B). En jämförelse mellan tabellerna visar att det endast är ANCOVA som resulterar i en signifikant behandlingseffekt; inkluderandet av Matteresultat vid förmätningen som kovariat minskar variationen inom grupperna (jfr de två Mean Square-värdena för Error i tabell A och B).

ANOVA för upprepade mätningar

ANOVA för upprepade mätningar (eng. *Repeated measures ANOVA*) kan också användas när samma individ är mätt upprepade gånger och mätningarna därmed är beroende av varandra (figur 13:2, fråga 3). Frågeställningen vid en ANOVA för upprepade mätningar skiljer sig från frågeställningen för en ANCOVA där förmätningen används som kovariat (Knapp & Schafer, 2009). Vid ANCOVA handlar det om interventionseffekten på eftermätningen efter justering för förmätningsvärdet. Själva förändringsvärdena och om grupperna skiljer sig åt när det kommer till dessa värden undersöks däremot vid ANOVA för upprepade mätningar. Med hjälp av en ANOVA för upprepade mätningar blir det möjligt att kontrollera för variationen i förändring över tid som är orsakad av skillnader mellan individer. Samma antaganden som tidigare gjordes för en ANCOVA gäller även för en ANOVA för upprepade mätningar (Dimitrov & Rumrill, 2003).

En ANOVA för upprepade mätningar måste byggas på att samma utfallsmått används vid förmätning och eftermätning (t.ex. att samma mattetest eller bedömningsinstrument används) – till skillnad från ANCOVA där andra kontinuerliga sekundära orsaksvariabler (t.ex. poäng från ett annat bedömningsinstrument) också kan inkorporeras. En annan skillnad är hur analysen behandlar förmätningen. Som beskrivs ovan, finns det en kategorisk faktor (Behandling) av intresse i en ANCOVA – vilket skiljer sig från en ANOVA



Figur 13:13. Ett exempel på hur en ANOVA för upprepade mätningar-design är upplagt.

för upprepade mätningar, där två kategoriska faktorer skapas. I figur 13:13 ges ett exempel på en ANOVA för upprepade mätningar med samma datamaterial som använts i figur 13:12. Individerna 1 till 9 har fördelats till grupper samt mätts vid två tillfällen (T1 – förmätning och T2 – eftermätning).

Tabell A i figur 13:14 visar resultaten av ANOVA:n för upprepade mätningar. Tre *F*-värden rapporteras, varav de två så kallade inom-individer-effekterna (eng. *within subjects effects*) är av intresse vid tolkning av resultaten i en effektutvärdering. Däremot är effekten av ”Behandling” (den så kallade mellan-individer-effekten, eng. *between subjects effect*) vanligtvis inte av intresse i en effektutvärdering (Dimitrov & Rumrill, 2003; Huck & McLean, 1975; Jennings, 1988). I beräkningen av denna effekt läggs för- och eftermätningarna ihop och ett medelvärde för varje individ tas fram. Då testas om dessa medelvärden skiljer sig åt mellan grupperna (Hassmén & Koivula, 1996). Det är därför rimligt att behandlingseffekten vanligtvis inte är av intresse eftersom vi sällan bryr oss om individernas genomsnittliga värden över en tidsperiod, utan i stället fokuserar på skillnader mellan grupperna; antingen efter interventionen (som i ANCOVA) eller i förändringsvärdena (ANOVA för upprepade mätningar).

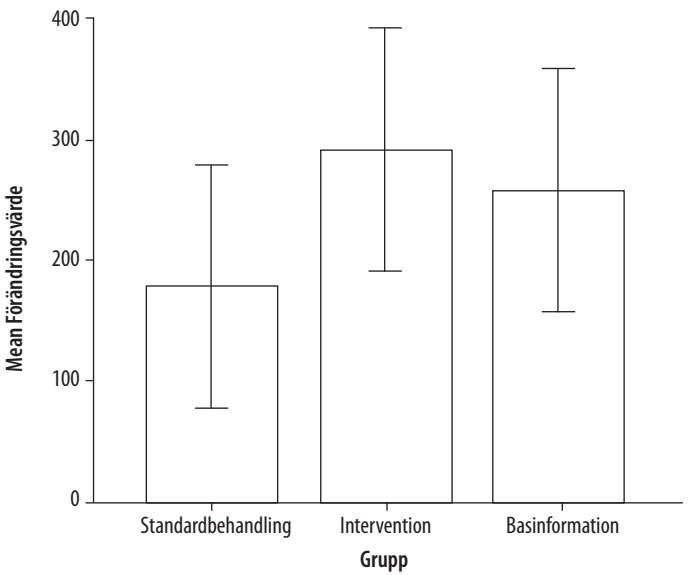
Inom-individer-effekten ”Tid” kan vara intressant om den visar att det generellt, oberoende av grupp, finns en statistiskt signifikant förändring. I vårt exempel visar resultaten på en ökad tendens i mattekunskap oavsett vilken grupp individerna tillhör (figur 13:14). I effektutvärderingar ligger fokus dock på inom-individer-effekten

A

	Mean Square	d.f.	F	Sig.
Between subjects effects				
Behandling	4 728	2	0,14	0,873
Error	34 077	6		
Within subjects effects				
Tid	235 984	1	205,39	0,000
Tid x Behandling	**4 720**	**2**	**4,11**	**0,075**
Error	1 149	6		

B

	Mean Square	d.f.	F	Sig.
Behandling	9 440	2	**4,11**	**0,075**
Error	2 298	6		



Figur 13:14. ANOVA för upprepade mätningar för Matteresultat vid för- respektive eftermätning (A). Envägs-ANOVA med Förändringsvärde som utfallsvariabel (B).

”Tid x Behandling”. Interaktionseffekten i tabell A (figur 13:14) är inte statistiskt signifikant ($p > 0{,}05$), vilket betyder att grupperna inte skiljer sig åt när det gäller förändring i Matteresultat mellan för- och eftermätning.

Resultatet som förklaras av interaktionseffekten i en ANOVA för upprepade mätningar kan också nås med hjälp av en envägs-ANOVA av förändringsvärden (eng. *change scores*) (Dimitrov & Rumrill,

2003; Huck & McLean, 1975). Förändringsvärden måste först räknas ut manuellt genom att subtrahera förmätningsvärdet från eftermätningsvärdet för varje individ. Analys av förändringsvärden i vårt exempel[6] visar att en envägs-ANOVA med Förändringsvärde som utfallsvariabeln ger samma *F*-värde och *p*-värde för behandlingseffekten (tabell B) som interaktionen mellan Tid och Behandling i en ANOVA för upprepade mätningar ger (tabell A) (figur 13:14).

I figur 13:14 åskådliggörs orsaken till att behandlingseffekten i analysen av förändringsvärden inte är signifikant; medelvärdena mellan grupperna ligger inom två standardavvikelser från varandra (d.v.s. linjerna överlappar staplarna i figur 13:14), vilket betyder att skillnaden mellan grupperna inte är signifikant. Med andra ord är variationen i förändringsvärden inom grupperna större än variationen mellan grupperna. I detta exempel hade individer med lägre matteresultat vid förmätning större förändringsvärden än individer med högre matteresultat vid förmätning, oavsett vilken grupp de tillhörde – vilket antyder att individer med sämre utgångsläge generellt sett har större förbättringspotential oavsett vilken grupp de tillhör.

Det här icke-signifikanta resultatet av envägs-ANOVA av förändringsvärden skiljer sig från den signifikanta behandlingseffekten i en ANCOVA med förmätningsvärdet som kovariat (figur 13:12). Anledningen till att resultaten oftast skiljer sig åt är att frågeställningarna som dessa två typer av analys bygger på är olika (Dimitrov & Rumrill, 2003; Knapp & Schafer, 2009). I ANCOVA används förmätningen för att *justera utfallsvariabeln* inför ett test av skillnaderna mellan grupperna vid eftermätningen. I en ANOVA baserad på förändringsvärde inkluderas däremot förmätningen för att undersöka interventionseffekten på *förändring* mellan för- och eftermätning. ANOVA för upprepade mätningar används även för att utforska förändring över flera tidspunkter för att undersöka om en behandlingseffekt kvarstår, stärks eller försvagas över tid. Det finns

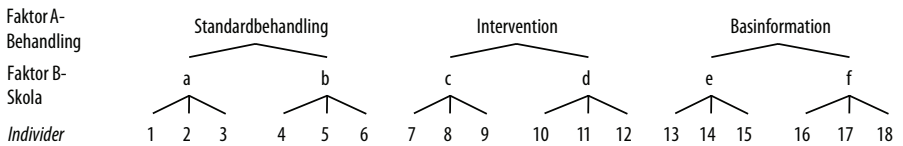
6 Se bilaga 1D för datamaterialet till exemplet som inkluderar förändringsvärden.

många situationer där det är relevant att utforska båda dessa frågeställningar och det är därför helt rimligt att utföra både en ANCOVA och en ANOVA för upprepade mätningar. Båda resultaten kan då användas för att få bättre kunskap om orsakerna bakom de observerade resultaten.

En ANOVA för upprepade mätningar är känslig för bortfall av individer. Om ett mättillfälle saknas exkluderar statistikprogrammet individen automatiskt från analysen. Ett annat problem är när det finns tre eller flera mättillfällen och spridningen av förändringsvärdena skiljer sig åt mellan de olika mättillfällena (t.ex. när spridningen av förändringsvärdena från eftermätning till sexmånadersuppföljning är större än spridningen av förändringsvärdena från sexmånadersuppföljning till ettårsuppföljning) (Brown m.fl., 2008; Quinn & Keough, 2002). Skillnader mellan individer i förändringsmönstret över tid kan inte heller utforskas med en ANOVA för upprepade mätningar (eftersom man bara testar om grupperna skiljer sig åt i deras genomsnittliga förändring). Därför rekommenderas övervägande av andra analystyper för att utforska förändring över tid, exempelvis multivariat variansanalys (eng. *Multivariate analysis of variance*, MANOVA), *Latent Growth Modeling (LGM)* eller *Hierarchical modeling (HM,* också kallat *multilevel models)* (se Brown m.fl., 2008).

Nästad ANOVA

Nästad (eng. *nested*) ANOVA är lämplig att använda när randomisering är på gruppnivå i stället för individnivå. Randomisering på gruppnivå kan vara praktiskt när en intervention ges till befintliga grupper (t.ex. en hel skolklass). Mätningarna blir därigenom beroende av varandra (figur 13:2, fråga 3) och risken är att jämförelserna mellan grupperna blir felaktiga och att en sann nollhypotes därigenom förkastas (typ I-fel). Beräkningen av *F*-värdet i en nästad ANOVA tar hänsyn till sambandet mellan individerna inom grupperna. Därigenom minskas även ovidkommande variation som beror på skillnader mellan exempelvis skolklasser.



Figur 13:15. Nästad ANOVA-design med tre nivåer (Faktor A, Faktor B och individer).

Figur 13:15 visar hur en sådan design kan se ut. Den första nivån (Faktor A: Behandling) består av tre grupper. Den underliggande nivån (Faktor B: Skola) består av sex skolor, med två skolor i varje grupp (d.v.s. två skolor är ”nästade” inom varje grupp). Effekten av faktor B är därför representerad i ANOVA-tabeller som ”Faktor B (Faktor A)”, det vill säga Faktor B nästad inom Faktor A. Varje skola är unik för just sin respektive grupp, vilket innebär att sex olika skolor ingår i studien (d.v.s. skolor a till f). Tre individer från varje skola har fått interventionen. En nästad design skiljer sig från en korsad faktoriell design med två faktorer (se figur 13:8) eftersom alla skolor inte är representerade i varje grupp.

Den nästade faktorn på den andra nivån (Faktor B, som i detta exempel är Skola) brukar vara en *random*-faktor (slumpfaktor). Ett kännetecken för en random-faktor är att om samma effektutvärdering utfördes igen skulle samma skolor inte nödvändigtvis väljas eftersom de inkluderats för att generalisera resultaten till en större population (t.ex. elever i alla skolor i Sverige) (Brown m.fl., 2008). I en effektutvärdering är en jämförelse av skillnader mellan enheterna (i det här fallet skolorna) på den här nivån sällan av intresse, i stället inkluderas random-faktorn för att kontrollera för ovidkommande variation. Om intresset i stället är att utforska effekten på endast dessa utvalda skolor (d.v.s. man generaliserar inte till en större population), blir den nästade faktorn en *fixed*-faktor (fast faktor). I analyserna som beskrivits hittills har alla orsaksvariabler av intresse varit fixed-faktorer eftersom intresset inte var att jämföra vilka grupper eller till exempel utbildningsnivåer som helst. Faktor A (d.v.s. Behandling) är därför alltid en fixed-faktor i en effektutvärdering.

Det är viktigt att förstå skillnaden mellan fixed-faktorer och ran-

dom-faktorer och att vara noggrann med att kontrollera om de är definierade som fixed eller random i det statistikprogram som används (vissa program utgår från att alla faktorer är fixed). I en nästad ANOVA har definitionen av faktorerna som fixed eller random avgörande konsekvenser för hur *F*-värdet beräknas för Faktor A. Som framgår tidigare är *F*-värdet:

$$F = \frac{\textit{variationen mellan grupperna}}{\textit{variationen inom grupperna}}$$

När Faktor A (Behandling) är fixed och Faktor B (Skola i vårt exempel) är en *random*-faktor, vilket är det vanliga fallet, blir *F*-värdet för Faktor A (d.v.s. behandlingseffekten):

$$F = \frac{\textit{variationen mellan grupperna}}{\textit{variationen mellan } \textbf{\textit{skolorna}} \textit{ inom grupperna}}$$

som är:

$$F = \frac{\textit{Mean Square-värdet för Faktor A}}{\textit{Mean Square-värdet för Faktor B inom A (B(A))}}$$

I kontrast, när både Faktor A (Behandling) och Faktor B (t.ex. de ”specifika” skolorna av intresse) är *fixed*, blir *F*-värdet för Faktor A (d.v.s. behandlingseffekten):

$$F = \frac{\textit{variationen mellan grupperna}}{\textit{variationen mellan } \textbf{\textit{individerna}} \textit{ inom skolorna}}$$

som är:

$$F = \frac{\textit{Mean Square-värdet för Faktor A}}{\textit{Mean Square-värdet för Error}}$$

Beskrivningen av *F*-värdena visar hur en nästad ANOVA tar hänsyn till beroende mellan mätningar. När den nästade faktorn, i det här fallet Skola, är en random-faktor (vilket är vanligast i effektutvärderingar), jämförs variationen mellan grupperna med variationen mellan skolorna. Med andra ord betyder detta att variationen inom

grupperna räknas ut genom att titta på hur mycket *skolornas* medelvärden skiljer sig åt från deras respektive grupps medelvärde. Detta skiljer sig från de variansanalyser som beskrivits hittills där variation inom grupperna räknas ut genom att titta på hur mycket *individernas* värden skiljer sig från deras respektive grupps medelvärde.

Tabell A i Figur 13:16 visar resultaten av en nästad ANOVA med Skola som en nästad random-faktor[7]. Resultaten visar ingen effekt av interventionen (d.v.s. effekten av Behandling är inte statistiskt signifikant) medan det finns signifikanta skillnader mellan skolorna inom grupperna (d.v.s. effekten av Skola [Behandling] är statistiskt signifikant). Att Skola är en random-faktor syns eftersom *F*-värdet för behandlingseffekten är Mean Square-värdet för Behandling dividerad med Mean Square-värdet för Skola (Behandling) (d.v.s. Behandling $F = 52774/58756 = 0{,}90$). En envägs-ANOVA utan Skola som nästad faktor visar hur viktigt det är att kontrollera för systematiska skillnader på grund av grupprandomisering; utan Skola som en random-faktor skulle den felaktiga slutsatsen att det finns en effekt av interventionen ha dragits (tabell B). Staplarna i figur 13:16 åskådliggör att detta är för att skillnaderna mellan skolorna inom grupperna är större än skillnaderna mellan grupperna.

Nästad ANOVA kan utökas till att inkludera flera nivåer av nästade och korsade faktorer (t.ex. Behandling korsad inom Skola och Skolklass nästad inom Behandling, med flera individer i varje klass). Om man vill öka sannolikheten att upptäcka en sann skillnad mellan grupperna på en viss nivå (d.v.s. öka den statistiska styrkan) i en nästad design ska man inte öka antalet individer utan man måste i stället öka antalet enheter (t.ex. skolklasser eller skolor) i den underliggande nivån.

Nästad ANOVA har utvecklats för experimentella designer. Detta betyder att nästad ANOVA är känslig för obalanserade designer, det vill säga när individer saknas i celler. Detta kan till exempel uppstå vid randomisering till grupperna på skolklassnivån, när skolorna är

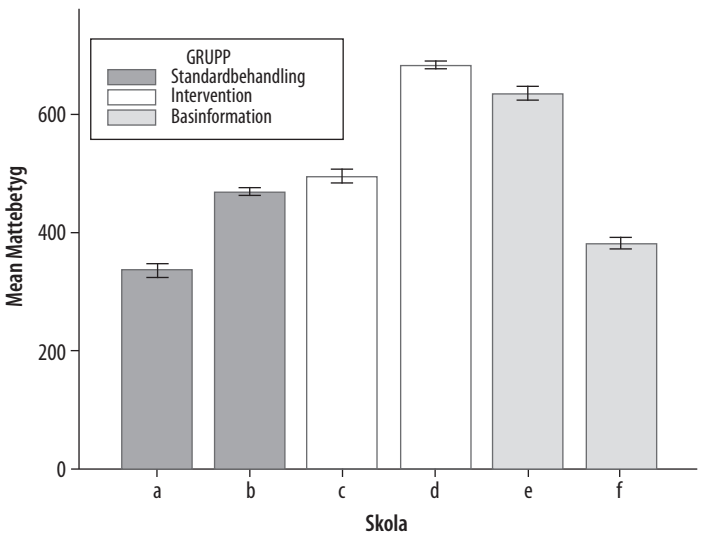
7 Se bilaga 1E för datamaterial till exemplet.

A

	Mean Square	d.f.	F	Sig.
Behandling	52 774	2	0,90	0,495
Skola (Behandling)	58 756	3	2835,44	0,000
Error	21	12		

B

	Mean Square	d.f.	F	Sig.
Behandling	52 774	2	4,49	0,03
Error	11 768	15		



Figur 13:16. Nästad ANOVA med Skola som nästad random faktor (A), envägs-ANOVA med samma datamaterial men utan faktorn Skola (B).

olika stora och när antalet skolklasser inte räcker till. En av anledningarna till att multilevel-regression utvecklades var för att hantera just sådana fall av obalanserade data (se Brown m.fl., 2008; Snijders & Bosker, 2012).

Variansanalys av mer komplicerade experimentella designer

De typer av variansanalys som beskrivs i kapitlet kan kombineras och utökas genom att inkludera flera kategoriska variabler (faktorer), kontinuerliga variabler (kovariater), upprepade mätningar och nästade faktorer av intresse. Om till exempel skillnader mellan indi-

vider vid studiens start kontrolleras för genom att inkludera förmätningen i en ANCOVA, erhålls ingen kunskap om hur andra faktorer påverkar effekten av interventionen. Därför är det ofta önskvärt att inte bara inkludera en förmätning i en ANCOVA utan också andra faktorer av intresse. Ett annat exempel är att inkludera flera sekundära kovariater och faktorer i en ANOVA för upprepade mätningar. Detta för att utforska hur de samvarierar med interventionen och påverkar ändring i utfallet över tid. Sådana mer komplicerade designer kräver noggranna överväganden av den experimentella designen och antalet individer eller enheter i den nästade nivån som behövs för att uppnå ett rimligt statistiskt styrkevärde.

Variansanalys kan även utökas till att inkludera flera utfallsvariabler av intresse och blir då en multivariat variansanalys (eng. *Multivariate analysis of variance;* MANOVA). En MANOVA kan vara tillämplig för att utforska interventionseffekten med hjälp av flera begreppsligt parallella mått. En MANOVA är att föredra framför flera separata variansanalyser för varje enskilt mått eftersom risken att förkasta en sann nollhypotes (typ I-fel) ökar med antalet tester (Harris, 1993).

Sammanfattning

Variansanalys är särskilt lämpad för att jämföra grupper av data, till exempel en interventionsgrupp och en jämförelsegrupp. Genom en välplanerad experimentell design och val av rätt variansanalys blir det möjligt både att kontrollera för ovidkommande variation i en effektutvärdering och att utforska de interagerande effekterna av flera sekundära orsaksvariabler av intresse. En variansanalys börjar alltid med att utforska om utfallsvariabeln uppfyller antaganden om normalfördelning och homogena varianser. Om så inte är fallet kan en datatransformation vara en lösning. Om utfallsvariabeln av primärt intresse är kategorisk kan variansanalys användas. Ovidkommande variation kontrolleras för och effekterna av andra variabler utforskas genom inkluderandet av sekundära orsaksvariabler av in-

tresse. Kategoriska och kontinuerliga orsaksvariabler analyseras med flervägs-ANOVA respektive ANCOVA. Variation mellan individer kan minskas med hjälp av en förmätning som kovariat i en ANCOVA. Upprepade mätningar kan också användas för att utforska förändring i utfallet över tid genom en ANOVA för upprepade mätningar. Nästad ANOVA möjliggör jämförelser av grupperna efter statistisk kontroll för beroende mellan mätningar på grund av randomisering på gruppnivån.

Det bör betonas att experimentell kontroll alltid är mer effektiv än statistisk kontroll och att det i vissa fall inte finns något sätt att kontrollera för oönskad variation på grund av ovidkommande faktorer. Det finns flera antaganden som datamaterialet måste uppfylla inför variansanalys, men effekten av avvikelser från dessa antaganden är små i jämförelse med metodiska fel i design och genomförande. Det viktigaste rådet är att noga tänka på en effektutvärderings frågeställningar, design och analys före datainsamlingen. I vissa fall kan detta betyda att variansanalys inte är lämplig. En förståelse av de olika typerna av variansanalyser ger inte bara statistiska kunskaper utan även en logisk utgångspunkt inför planeringen av experimentell design.

Fördjupningslitteratur

Borg, E. & Westerlund, J. (2010). *Statistik för beteendevetare.* Andra upplagan. Stockholm: Liber.

Brown, C. H., Costigan, T. E. & Kendziora, K. T. (2008) Data analytic frameworks: Analysis of variance, latent growth, and hierarchical models. In Nezu, A. M. & Nezu, C. M. (Eds.) *Evidence-based outcome research: A practical guide to conducting randomized controlled trials for psychosocial interventions*. New York: Oxford University Press.

Hassmén, P. & Koivula, N. (1996). *Variansanalys.* Lund: Studentlitteratur.

Miller, G.A. & Chapman, J.P. (2001). Misunderstanding Analysis of Covariance. *Journal of Abnormal Psychology*. 110(1), 40-48.

Scheffé, H. (1999). *The analysis of variance.* New York: Wiley.

Referenser

Borg, E. & Westerlund, J. (2010). *Statistik för beteendevetare. Andra upplagan.* Stockholm: Liber.

Brown, C. H., Costigan, T. E. & Kendziora, K. T. (2008). Data analytic frameworks: Analysis of variance, latent growth, and hierarchical models. I A. M. Nezu & C. M. Nezu (Eds.), *Evidence-based outcome research: A practical guide to conducting randomized controlled trials for psychosocial interventions* (pp. 285-313). New York: Oxford University Press.

Coombs, W. T., Algina, J. & Oltman, D. O. (1996). Univariate and omnibus hypothesis tests selected to control Type I error rates when population variances are not necessarily equal. *Review of Educational Research, 66*, 137–179.

Dimitrov, D. M. & Rumrill, P. D. J. (2003). Pretest-posttest designs and measurement of change. *Work, 20*, 159–165.

Dobson, A. J. (2001). *An introduction to Generalized Linear Models* (2nd ed.). London: Chapman & Hall.

Guo, S. & Fraser, M. W. (2010). *Propensity Score Analysis*. California: SAGE Publications.

Hancock, G. R. & Klockars, A. J. (1996). The quest for alpha: developments in multiple comparison procedures in the quarter century since Games (1971). *Review of Educational Research, 66*, 269–306.

Harris, R. J. (1993). Multivariate analysis of variance. In L. K. Edwards (Red.), *Applied Analysis of Variance in Behavioral Science* (s. 691–716). New York: Marcel Dekker.

Hassmén, P. & Koivula, N. (1996). *Variansanalys*. Lund: Studentlitteratur.

Hochberg, Y. & Tamhane, A. C. (1987). *Multiple Comparison Procedures*. New York: Wiley.

Huberty, C. J. (1993). Historical origins of statistical testing practices: the treatment of Fisher versus Neyman-Pearson views in textbooks. *Journal of Experimental Education, 61*, 317–333.

Huck, S. W. & McLean, R. A. (1975). Using a repeated measures ANOVA to analyze data from a pretest-posttest design: A potentially confusing task. *Psychological Bulletin, 82*, 511–518.

Jamieson, J. (1999). Dealing with baseline differences: two principles and two dilemmas. *International Journal of Psychophysiology, 31*(2), 155-161. doi: 10.1016/s0167-8760(98)00048-8

Jennings, E. (1988). Models for pretest-posttest data: repeated measures ANOVA revisited. *Journal of Educational Studies, 13*, 273–280.

Kirk, R. E. (1995). *Experimental Design*. Pacific Grove: Brooks/Cole.

Knapp, T. R. & Schafer, W. D. (2009). From Gain Score to ANCOVA F (and vice versa). *Practical Assessment, Research & Evaluation, 14*(6), 1–7.

Langsrud, Ø. (2003). ANOVA for unbalanced data: Use Type II instead of Type III sums of squares. *Statistics and Computing, 13*(2), 163–167. doi: 10.1023/a:1023260610025

Maxwell, S. E. & Delaney, H. D. (1990). *Designing experiments and analyzing data: a model comparison perspective*. Belmont, California: Wadsworth Publishing.
McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed.). London: Chapman & Hall.
Miller, G. A. & Chapman, J. P. (2001). Misunderstanding Analysis of Covariance. *Journal of Abnormal Psychology, 110*(1), 40–48.
Milliken, G. A. & Johnson, D. E. (1984). *Analysis of messy data. Vol. 1: designed experiments*. New York: Van Nostrand Reinhold.
Quinn, G. P. & Keough, M. J. (2002). *Experimental Design and Data Analysis for Biologists*. Cambridge: Cambridge University Press.
Senn, S. (2006). Change from baseline and analysis of covariance revisited. *Statistics in Medicine, 25*, 4334–4344.
Snijders, T. A. B. & Bosker, R. J. (2012). *Multilevel Analysis: An introduction to basic and advanced multilevel modeling* (2 ed.). London: Sage Publications.

Bilaga 1

Tabell A. Datamaterial som analyseras med ett *t*-test med Matteresultat som utfallsvariabel och Behandling som orsaksvariabel.

Individ	Behandling (Grupp)	Matteresultat
1	Standardbehandling	350
2	Standardbehandling	325
3	Standardbehandling	330
4	Intervention	450
5	Intervention	425
6	Intervention	430

Tabell B. Datamaterial som analyseras med en envägs-ANOVA med Matteresultat som utfallsvariabel och Behandling som orsaksvariabel. En ANCOVA möjliggör inklusion av kovariatet Föräldrars inkomst.

Individ	Behandling (Grupp)	Matteresultat	Föräldrars inkomst
1	Standardbehandling	340	390
2	Standardbehandling	335	357
3	Standardbehandling	330	333
4	Intervention	440	387
5	Intervention	435	363
6	Intervention	430	330
7	Basinformation	400	393
8	Basinformation	395	360
9	Basinformation	390	333

Tabell C. Datamaterial som analyseras med en tvåvägs-ANOVA med Matteresultat som utfallsvariabel och Behandling och Föräldrars utbildning som orsakvariabler.

Individ	Behandling (Grupp)	Matteresultat	Föräldrars utbildning
1	Standardbehandling	330	Grundskola
2	Standardbehandling	331	Grundskola
3	Standardbehandling	335	Gymnasium
4	Standardbehandling	338	Gymnasium
5	Standardbehandling	424	Högskola
6	Standardbehandling	438	Högskola
7	Intervention	439	Grundskola
8	Intervention	451	Grundskola
9	Intervention	448	Gymnasium
10	Intervention	443	Gymnasium
11	Intervention	439	Högskola
12	Intervention	445	Högskola
13	Basinformation	395	Grundskola
14	Basinformation	393	Grundskola
15	Basinformation	391	Gymnasium
16	Basinformation	398	Gymnasium
17	Basinformation	434	Högskola
18	Basinformation	438	Högskola

Tabell D. Datamaterial som analyseras med en ANCOVA med Matteresultat vid eftermätning som utfallsvariabeln, och Behandling och Matteresultat vid förmätning som orsakvariabler. Alternativt analyseras datamaterialet med en ANOVA för upprepade mätningar eller en envägs-ANOVA med Förändringsvärde som utfallsvariabel och Behandling som orsaksvariabel.

Individ	Behandling (Grupp)	Matteresultat förmätning	Matteresultat eftermätning	Förändrings-värde
1	Standardbehandling	330	446	116
2	Standardbehandling	128	339	211
3	Standardbehandling	42	223	181
4	Intervention	333	565	232
5	Intervention	124	455	331
6	Intervention	43	322	279
7	Basinformation	331	517	186
8	Basinformation	126	400	274
9	Basinformation	44	295	251

Tabell E. Datamaterial som analyseras med en nästad ANOVA med Behandling och den nästade random-faktorn Skola som orsakvariabler.

Individ	Behandling (Grupp)	Matteresultat	Skola
1	Standardbehandling	340	a
2	Standardbehandling	335	a
3	Standardbehandling	330	a
4	Standardbehandling	464	b
5	Standardbehandling	469	b
6	Standardbehandling	471	b
7	Intervention	500	c
8	Intervention	495	c
9	Intervention	490	c
10	Intervention	682	d
11	Intervention	686	d
12	Intervention	678	d
13	Basinformation	640	e
14	Basinformation	635	e
15	Basinformation	630	e
16	Basinformation	380	f
17	Basinformation	376	f
18	Basinformation	385	f

Kapitel 14 | Maria Paola Caria & Maria Rosaria Galanti

Regressionsanalys[1]

Regressionsanalysens principer

Regressionsanalys är en statistisk metod för analys av sambandet mellan en beroendevariabel och en eller flera förklarande variabler. Beroendevariabeln benämns ofta utfalls- eller responsvariabel och betecknas vanligtvis med symbolen Y. Den förklarande variabeln benämns ofta exponeringsvariabel, kovariat eller prediktor och betecknas med symbolen X. Bas för regressionsanalysen är en regressionsekvation som beskriver sambandet mellan variablerna genom okända parametrar och som gör det möjligt att skatta en förändring på en beroendevariabel förutsatt förändring på en eller fler oberoende variabler. En regressionsekvation kan ritas upp som en linje i ett diagram.

I forskning om interventioner representerar beroendevariabeln alltid resultatets utfall och den centrala prediktionsvariabeln representerar alltid interventionen som utvärderas. Det huvudsakliga syftet med regressionsanalys i utvärderingsforskning är därför att få en uppskattning av den parameter som bäst beskriver förhållandet mellan interventionen och det utfall man vill studera. Om interventio-

1 Kapitlet har översatts från engelska till svenska av Åsa Kling, institutionen för psykologi, Uppsala universitet.

nen under utvärdering är den enda oberoende variabeln i analysen talar vi om *enkel* regression (eng. *simple regression*). Alternativt finns det fler oberoende variabler och då handlar det om *multipel* regression. Vi kan exempelvis utvärdera effekten av en intervention för att förebygga självmord och ta hänsyn till kön, ålder, utbildningsnivå, yrke och bostadsort utöver själva interventionen.

Typen av beroendevariabel styr vilken sorts regression som ska användas; om beroendevariabeln är kontinuerlig används linjär regression och om den är dikotom används logistisk regression. I båda fallen kan de oberoende variablerna vara antingen kontinuerliga, dikotoma eller kategoriska med fler än två nivåer.

Regressionsteknik kan användas i både experimentella studier och observationsstudier.

Linjär regression i effektutvärderingar

Innan vi förklarar de grundläggande principerna för regression, går vi igenom analysen av en kontinuerlig utfallsvariabel (d.v.s. linjär regressionsanalys). Linjär regression går ut på att skatta parametrar till en rät linje anpassad till en uppsättning observationer. Den matematiska formeln för den enkla linjära regressionsmodellen är:

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad \text{(modell 1)}$$

där *i* antar värden från 1 till antalet observationer i studien; y_i är utfallsvariabeln och x_i är värdet på interventionsvariabeln för den i:e observationen; interceptet, β_0, uppskattar medelvärdet för utfallsvariabeln när intervention är noll (d.v.s. ingen intervention) och linjens lutning, β_1, uppskattar genomsnittlig förändring i utfallsvariabeln för en enhets förändring i interventionsvariabeln. Om β_1 är positiv ökar medelvärdet av Y när interventionsvariabeln ökar och om β_1 är negativ minskar medelvärdet av Y när interventionsvariabeln ökar. Om $\beta_1 = 0$ har interventionen ingen effekt på Y. e_i är skillnaden mellan det faktiska värdet på utfallsvariabeln och det predicerade vär-

det för den i:e observationen. Den benämns ofta fel eller residual för att det är den del av y_i som inte förklaras av modellen. e_i antas vara en slumpmässig och normalfördelad avvikelse.

Generellt sett kan variansen hos observationerna delas in i varians som kan förklaras av de oberoende variablerna (modellvarians) och varians som inte förklaras av de oberoende variablerna (felvarians eller residual varians). Det är variansen av e_i, σ_e^2, som är ”felvariansen”. De två parametrarna β_0 och β_1 utgör modellens fixa del; felen och felens varians är modellens slumpmässiga del. Proportionen varians som förklaras av de oberoende variablerna uppskattas genom att dela modellvariansen med den totala variansen och kallas R^2.

Liksom i variansanalys är några av de grundläggande antagandena i linjär regressionsanalys:

- Oberoende: varje observation är oberoende av de andra observationerna.
- Homogen varians (”homoskedasticitet”): observationerna har samma varians eller jämn spridning kring regressionslinjen för alla värden på den oberoende variabeln.
- Normalfördelning: de observerade värdena av Y är normalfördelade kring varje predicerat värde på linjen. Det betyder att om upprepade mätningar av Y görs för ett visst värde på X förväntas de flesta av dessa observationer hamna nära regressionslinjen och bara några få hamna långt från linjen.

Ytterligare antaganden som är specifika för linjär regression är:

- Linearitet: det underliggande sambandet mellan den beroende och de oberoende variablerna följer en rät linje.
- Frånvaro av multikollinearitet: de oberoende variablerna får inte korrelera med varandra (d.v.s. inte vara lineärt relaterade). Det största problemet är att om graden av multikollinearitet ökar, blir koefficienterna i modellen instabila och deras standardfel kan öka dramatiskt. Om vi exempelvis använder de fyra variablerna mammas utbildning, pappas utbildning, mammas inkomst och pappas inkomst som indikatorer i en modell för att uppskatta fa-

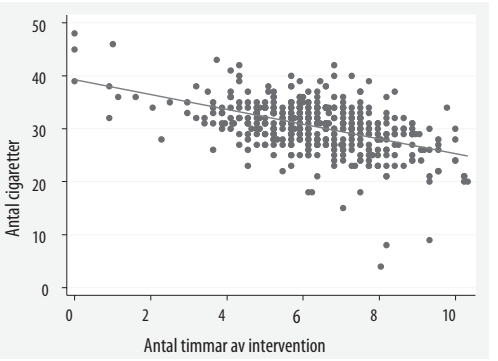
miljens socioekonomiska status, kommer några av dessa variabler att vara kraftigt korrelerade. Grad av kollinearitet kan kontrolleras med hjälp av ett toleransindex. Ett toleransvärde på 0,1 betyder att bara 10 procent av en variabels varians inte förklaras av andra prediktorerna i modellen. En tumregel är att om en variabels toleransvärde är lägre än 0,1 ska den tas bort från modellen.

Exempel 1

Anta att vi i en studie ska utvärdera effekten av ett preventivt hälsoutbildningsprogram i skolan mot rökning bland ungdomar. Specifikt vill vi ta reda på sambandet mellan hur många timmar ungdomarna deltagit i utbildningen (X) och hur många cigaretter de rökt under månaden efter interventionen (Y). Figur 14:1 visar resultaten av den enkla regressionsanalysen och ekvation 1 beskriver den linjära regressionsmodellen.

$$\text{Antal cigaretter}_i = \beta_0 + \beta_1 \text{ antal utbildningstimmar}_i + e_i \qquad \text{(modell 2)}$$

Modellen ger de skattade värdena (för linjens anpassning till data) $\beta_0 = 40$ (standardavvikelse=0,22) och $\beta_1 = -1{,}6$ (standardavvikelse=0,04) med felvarians $\sigma_e^2 = 12{,}7$. I vårt exempel är y_i och x_i antal rökta cigaretter respektive antal utbildningstimmar för den i:e ung-



Figur 14:1. Regressionslinje för sambandet mellan antalet rökta cigaretter under månaden efter interventionen och antal utbildningstimmar.

domen. e_i är skillnaden mellan det faktiska antalet och det predicerade antalet rökta cigaretter för den i:e ungdomen.

Tolkningen av koefficienterna β_0 och β_1 är lätt att se. Interceptet β_0 uppskattar genomsnittligt antal rökta cigaretter under månaden efter interventionen till 40 för de ungdomar som inte fick någon utbildning (d.v.s. 0 timmar). Regressionskoefficienten β_1 uppskattar att antalet rökta cigaretter minskar med i genomsnitt 1,6 för varje timmes utbildning.

Värdet inom parentes (0,04) är regressionskoefficientens standardavvikelse, och den kan användas för att bedöma signifikansen för sambandet mellan intervention och antal rökta cigaretter. Det görs genom att dela regressionskoefficienten med dess standardavvikelse. Den kvadrerade kvoten kallas Walds test och följer en chi²-fördelning med en frihetsgrad. I detta exempel blir Walds test $(-1{,}6/0{,}04)^2 = 1555$, vilket är starkt signifikant. Standardavvikelsen kan även användas för att skapa ett 95-procentigt konfidensintervall (KI) genom att ta regressionskoefficienten ±1,96 multiplicerat med standardavvikelsen. I detta exempel hamnar regressionskoefficientens konfidensintervall mellan gränserna -1,72 och -1,56, vilket inte inkluderar värdet 0. Både Walds test och konfidensintervallet på 95 procent tyder således på att interventionen lyckades med att få ungdomar att röka mindre.

Exempel 2

Tolkningen håller även med en dikotom interventionsvariabel. Anta att vi den här gången är intresserade av effekten när interventionen prövas i två kategorier: kontrollgrupp (x = 0) eller interventionsgrupp (x = 1). I det fallet blir den linjära modellen:

$$\text{Antal cigaretter}_i = \beta_0 + \beta_1 \text{intervention}_i + e_i \qquad \text{(modell 3)}$$

Den modellen ger de skattade värdena $\beta_0 = 41$ (0,26) och $\beta_1 = -10{,}6$ (0,32) med felvarians $\sigma_e^2 = 16{,}2$. Genomsnittligt antal rökta cigaretter för prediktorns referensvärde, det vill säga för ungdomar i

kontrollgruppen (intervention = 0) är 41. Koefficienten β_1 tolkas återigen som förändring av antal rökta cigaretter för varje enhets förändring av prediktorn, vilket i vårt exempel betyder att jämföra ungdomar i interventionsgruppen med ungdomar i kontrollgruppen. Eftersom β_1 är negativ, förväntas ungdomarna i interventionsgruppen röka i genomsnitt 10,6 färre cigaretter än ungdomarna i kontrollgruppen. Walds test för regressionskoefficienten är $(-10{,}6/0{,}32)^2 = 1096$, vilket återigen är starkt signifikant. Konfidensintervallet på 95 procent har gränserna -11,27 till -10,01.

Situationen i vårt första exempel är mest typisk för observationsstudier, där individer på ett naturligt sätt exponeras för olika doser av interventionen (Jacobs, De Bourdeaudhuij, Thijs, Dendale & Claes, 2011). I experimentell forskning däremot randomiseras vanligtvis inte individer till interventionen på en kontinuerlig skala utan på en dikotom (antingen intervention eller kontrollgrupp). Därför är situationen i vårt andra exempel mer vanlig i experimentella studier.

Kategoriska prediktorer med två eller flera nivåer fungerar också bra i regressionsanalys. Men det är däremot inte lämpligt att låta dem ingå i modellen som om de var kontinuerliga variabler. Det beror på att siffrorna som representerar de olika nivåerna i kategoriska prediktorer bara är markörer utan någon numerisk betydelse. Har man sådana variabler är alternativet att låta dem representeras av indikatorer eller dummyvariabler i regressionsmodellen. Dessa variabler är dikotoma kvalitativa indikatorer som bara har två möjliga värden, 0 eller 1. Man skapar en sådan dummyvariabel för varje kategori av prediktorn, där varje observation tar värde 1 om den tillhör den kategorin, annars tar den värde 0. I regressionsmodellen kommer k-1 dummyvariabler att inkluderas, det vill säga en för varje kategori förutom referenskategori (referensnivå). Valet av referensnivå bör göras på naturliga/logiska grunder, möjligtvis är den kategori som har störst urvalsstorlek att föredra (Hardy, 1993). Nedan ges ett exempel på en sådan prediktor och hur den används i regressionsanalys.

Exempel 3

Anta att vi vill utvärdera effekten av vårt preventionsprogram, vilket implementeras i tre nivåer (kategorier): låg = 0, mellan = 1 och hög = 2. Vi väljer gruppen ”låg” som referenskategori och kodar om de andra kategorierna till dikotoma indikatorer: Implementering $^{\text{MELLAN}}$ = 1 om implementeringsnivån = 1, annars 0; Implementering $^{\text{HÖG}}$ = 1 om implementeringsnivån = 2, annars 0. Modellen för data blir då:

$$\text{Antal cigaretter}_i = \beta_0 + \beta_1 \text{ implementering}^{\text{MELLAN}}{}_i + \beta_2 \text{ implementering}^{\text{HÖG}}{}_i + e_i \qquad \text{(modell 4)}$$

Det ger de skattade värdena β_0 = 41 (0,25), β_1 = -9,4 (0,34), och β_2 = -12,4 (0,36) med felvarians σ_e^2 = 14,2. Förväntat antal rökta cigaretter i prediktorns referensgrupp, alltså för de ungdomar som fick den låga nivån av interventionen, är 41. Ungdomar som fick interventionen på mellannivå förväntas röka i genomsnitt 9,4 cigaretter färre än ungdomar i referensgruppen. Ungdomar som fick den höga nivån av interventionen förväntas röka i genomsnitt 12,4 cigaretter färre än ungdomar i referensgruppen. Skillnaden mellan β_2 och β_1 = -3,0 jämför ungdomar som fick den höga implementeringen med ungdomar som fick mellannivån av implementeringen. Regressionsanalys med en kategorisk prediktor är ett sätt att jämföra flera grupper som motsvarar den analys som görs i variansanalys (ANOVA).

I många fall behöver vi ta med andra förklarande variabler utöver interventionen. Det är exempelvis inte ovanligt att vi vill undersöka samtidiga effekter av fler prediktorer på en utfallsvariabel. Dessutom måste vi justera analyserna för ovidkommande faktorer, så kallade störfaktorer (eng. *confounders*). Slutligen behöver vi ibland veta om de övriga prediktorerna som påverkar utfallet är oberoende av varandra eller om de interagerar. Multipel regressionsanalys är ett kraftfullt verktyg som gör det möjligt att undersöka många förklaringsvariabler på samma gång.

De statistiska analysmodellerna ANOVA och ANCOVA ger sam-

ma resultat som enkel respektive multipel regression förutsatt att exponeringsvariabeln, det vill säga interventionen, mäts som en kategorisk variabel. Om interventionen mäts som en kontinuerlig variabel kan man fortfarande använda ANOVA eller ANCOVA, men modellerna är då inte lika effektiva som regressionsmodeller. Regressionsanalys är även mer flexibel när man har utfallsvariabler som inte är kontinuerliga (se logistisk regression).

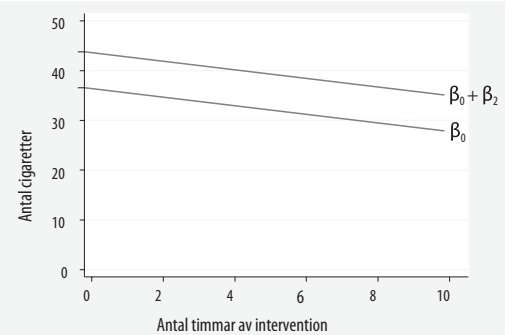
Exempel 4

Det är rimligt att anta att både omfattningen av en intervention och föräldrars rökning har inverkan på hur mycket ungdomar röker. För att undersöka dessa båda influenser på utfallsvariabeln infogas de i modellen som prediktorer. Föräldrars rökning kategoriseras som: ingen förälder röker (föräldrars rökning = 0) och minst en förälder röker (föräldrars rökning = 1). I modellen representeras interventionen av en kontinuerlig variabel för antal utbildningstimmar och föräldrars rökning representeras av en dikotom variabel.

$$\text{Antal cigaretter}_i = \beta_0 + \beta_1 \text{ antal utbildningstimmar}_i + \beta_2 \text{ föräldrars rökning}_i + e_i \qquad \text{(modell 5)}$$

Det ger de skattade värdena $\beta_0 = 37$ (0,30), $\beta_1 = -1{,}1$ (0,05) och $\beta_2 = 4{,}7$ (0,32), med felvarians $\sigma_e^2 = 9{,}7$. Koefficienterna β_1 och β_2 uppskattar de samtidiga effekterna av antal utbildningstimmar och föräldrars rökning på antal rökta cigaretter. Båda är statistiskt signifikanta. Koefficienten β_0 visar interceptet för situationen ingen intervention för ungdomar vars föräldrar inte röker, medan $\beta_0 + \beta_2$ visar interceptet för situationen ingen intervention för ungdomar med minst en förälder som röker. När vi har korrigerat modellen för föräldrars rökning ser vi med andra ord att interceptet för regressionslinjen är olika för de två kategorierna (figur 14:2).

Om ungdomarna som får en mer intensiv intervention har föräldrar som röker i mindre omfattning än ungdomar som får en mindre intensiv intervention kan den observerade effekten bero på influ-



Figur 14:2. Regressionslinjer för relationen mellan antal rökta cigaretter under månaden efter interventionen och antal utbildningstimmar korrigerade för föräldrars rökning.

ensen av föräldrarnas rökning och inte interventionen. Det skulle vara en klassisk situation där störfaktorer (eng. *confounders*) påverkar utfallet, vilket kan inträffa även i randomiserade studier. Multipel regression löser frågan om det finns störfaktorer på ett enkelt sätt, det vill säga om interventionseffekten förändras när föräldrars rökning tas med i modellen. För att ta reda på det jämför vi koefficient β_1 i modell 2 (”råskattningen”) med koefficient β_1 i modell 5 (den ”justerade skattningen”). I vårt fall förändras koefficienten från -1,6 till -1,1, en ökning med 30 procent. Det här resultatet tyder på att interventionens positiva effekt på rökning hos ungdomar delvis är sammanblandad med föräldrars rökning. I ett sådant fall är den justerade skattningen en mer valid skattning än råskattningen.

Att kontrollera för störfaktorer (eng. *confounders*) är av avgörande betydelse för valida skattningar av samband med observationsstudier och för att kunna uttala sig om möjliga kausala relationer. Randomisering i experimentella studier ökar sannolikheten för att grupperna som får olika behandling i genomsnitt är lika varandra. Men randomisering är ingen garanti för att grupperna är fullkomligt jämförbara (Stanley, 2007). Därför är det viktigt att även i randomiserade kontrollerade studier (RCT) alltid kontrollera för möjliga störfaktorer. I multipel regressionsanalys kan alla typer av sådana variabler kontrolleras.

Exempel 5

Genom multipel regression kan vi avslutningsvis direkt testa eventuella interaktioner eller modererande effekter. Vi kan exempelvis undersöka om interventionen har olika effekt på ungdomar beroende på föräldrars rökning i modellen:

$$\text{Antal cigaretter}_i = \beta_0 + \beta_1 \text{intervention}_i + \beta_2 \text{föräldrars rökning}_i + \beta_3 \text{intervention*föräldrars rökning} + e_i \quad \text{(modell 6)}$$

där koefficienten β_3 är produkten av de två prediktorerna. Den statistiska signifikansen hos produkttermen visar om det finns en modifierande effekt av föräldrars rökning. Med andra ord visar β_3 hur effekten av interventionen ändras bland ungdomar vars föräldrar röker respektive inte röker. Om β_3 inte är statistiskt signifikant betyder det att effekten av interventionen är lika mellan ungdomar vars föräldrar röker respektive inte röker, samt att effekten av föräldrars rökning är densamma i interventions- och kontrollgrupp. I detta fall kan vi ta bort produkttermen från modellen. Om β_3 är statistiskt signifikant betyder det att effekten av interventionen på medelvärde av antal cigaretter är olika bland ungdomar som har föräldrar som röker och dem vars föräldrar inte röker, samt att effekten av föräldrars rökning är olika i interventions- och kontrollgrupp. I detta fall bör produkttermen stå kvar i modellen. Tolkningen av de övriga parametrarna i en modell med en interaktionsterm är lite annorlunda: β_1 skattar effekten av interventionen när föräldrarnas rökning kodas som ”0”, det vill säga bland ungdomar vars föräldrar inte röker; β_2 skattar effekten av föräldrars rökning när interventionen är =0, det vill säga bland kontroller; summan $\beta_1 + \beta_3$ ger en skattning av effekten av interventionen när föräldrarna röker; summan $\beta_2 + \beta_3$ ger en skattning av effekten av föräldrars rökning i interventionsgruppen.

Logistisk regression i effektutvärderingar

Logistisk regression undersöker relationen mellan en dikotom beroendevariabel och en eller flera oberoende variabler som kan vara kontinuerliga, dikotoma eller kategoriska med fler än två nivåer. I allmänhet betecknar det dikotoma utfallsvärdet 1 en ”händelse” (det resultat som vi är intresserade av), medan värdet 0 betecknar en ”icke-händelse”. Utfallet kan vara naturligt dikotomt (såsom en specifik diagnos) eller så kan det göras dikotomt för att kunna användas i analysen genom att man sätter ett gränsvärde för en kontinuerlig variabel (BMI kan t.ex. dikotomiseras i övervikt mot inte övervikt). En sådan omvandling innebär alltid förlust av statistisk power. Det finns två huvudskäl till varför man ska göra en kontinuerlig variabel dikotom. Det första är att en dikotom variabel kan ha större praktisk mening än en kontinuerlig (t.ex. när det finns gränsvärden för kliniska beslut). Det andra är om variabeln inte är normalfördelad.

Den matematiska formeln för *enkel* logistisk regression (med en oberoende variabel) är:

$$P(y_i=1|x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$$

där y_i och x_i är det dikotoma utfallet respektive interventionsvärde för den i:e observationen, $P(y_i=1) = p_i$ = P(”händelse”) är sannolikheten för händelsen och $P(y_i=1|x_i)$ är den förväntade sannolikheten av händelsen för ett visst värde på interventionen. I motsats till linjär regression har inte koefficienten i logistisk regression någon enkel omedelbar tolkning. Men användbarheten blir uppenbar om vi i stället för sannolikheten beräknar oddsen för händelsen:

$$\text{Odds}(y_i=1|x_i) = \frac{p_i}{1-p_i} = \frac{e^{\beta_0+\beta_1 x_i}/(1+e^{\beta_0+\beta_1 x_i})}{1/(1+e^{\beta_0+\beta_1 x_i})} = e^{\beta_0+\beta_1 x_i}$$

Den naturliga logaritmen av oddsen för händelsen är:

$$\ln\left[\frac{p_i}{1-p_i}\right] = \ln\, e^{\beta_0 + \beta_1 x_i} = \beta_0 + \beta_1 \bullet x_i = \text{logit}\,(p_i)$$

Notera likheten med den linjära regressionsmodellen. Skillnaden är att i stället för att anta en linjär relation mellan interventionen och utfallets medelvärde antar vi en linjär relation mellan interventionen och logaritmen av utfallets odds (även kallad ”sannolikhetens logit”). Det kan dessutom visas att $e^{\beta_0}/(1 + e^{\beta_0})$ är händelsens sannolikhet när ingen intervention ges (intervention = 0) och e^{β_1} uppskattar oddskvoten för en enhets förändring i variabeln för intervention (intervention = 1).

Som metod kommer logistisk regression undan många av de förutsättningar som begränsar linjär regression, i synnerhet angående linjäritet, homogen varians och normalfördelning. Däremot finns det andra antaganden. I logistisk regression måste varje observation vara oberoende. Det får inte heller finnas hög eller helst någon multikollinearitet mellan kovariater. Logistisk regression kräver inte att de beroende och oberoende variablerna är linjärt relaterade men däremot att de oberoende variablerna är linjärt relaterade till log-oddsen.

Exempel 6

Anta att vi är intresserade av att undersöka om en preventiv intervention mot användning av alkohol har någon inverkan på förekomst av berusning under månaden efter interventionen. Utfallsvariabeln är nu dikotom (vara berusad eller inte). Vi låter y = 1 stå för att berusning har förekommit och y = 0 stå för att berusning inte har förekommit. Modellen är *enkel* logistisk regression, där intervention definieras i kategorierna kontrollgrupp (x = 0) eller interventionsgrupp (x = 1):

$$\text{logit}(p_i) = \beta_0 + \beta_1 \text{intervention}_i \qquad \text{(modell 7)}$$

Modellen ger de skattade värdena β_0 = -1,6 (0,09) och β_1 = -0,6 (0,16). Sannolikheten för rapporterad berusning för kovariater lika med 0, det vill säga för deltagare i kontrollgruppen, är $e^{\beta_0}/ (1+e^{\beta_0})$ = 0,17 = 17 %. Koefficienten e^{β_1} = 0,5 visar oddskvoten för jämförelsen mellan interventions- (x = 1) och kontrollgrupp (x = 0). Detta tolkas som att sannolikheten i interventionsgruppen för att rapportera förekomst av berusning åtminstone en gång är 0,5 gånger den för ungdomarna i kontrollgruppen. Ett 95-procentigt konfidensintervall kring oddskvoten kan beräknas genom $e^{\beta_1 \pm 1.96 \cdot SE}$, där SE är standardfelet och intervallgränserna blir 0,40 till 0,74. Eftersom värdet 1, som betyder att det inte finns något samband, inte ligger inom det intervallet, skiljer sig oddskvoten signifikant från 1. Regressionskoefficientens signifikans (d.v.s. oddskvotens signifikans) får man från p-värdena i Walds test $(-0,6 / 0,16)^2 = 14,1$, som är starkt signifikant. Både det konfidensintervallet på 95 procent och Walds test visar alltså att interventionen har en statistiskt signifikant preventiv effekt på förekomst av berusning hos ungdomar månaden efter interventionen.

Sammanfattningsvis är skillnaden mellan logistisk och linjär regression att utfallsvariabeln i logistisk regression är dikotom. Utöver den skillnaden gäller samma generella principer som för linjär regression även i logistisk regression. På motsvarande sätt kan fler oberoende variabler analyseras på samma gång i multipel logistisk regression, det är möjligt att kontrollera för störfaktorer och analyserna kan undersöka modifierande effekter (interaktioner).

Flernivåanalys i effektutvärderingar

Flernivåanalys (eng. *multilevel analysis*) kallas också hierarkiska modeller eller analyser med mixade effekter eller slumpeffekter. I de flesta fall syftar alla dessa olika termer på samma sak. Flernivådata uppkommer när det finns en hierarkisk struktur eller klusterstruk-

tur i data. Flernivåanalys började först användas inom skolforskning (Goldstein, 1987) där forskare insåg att observationer från elever i samma skola inte var oberoende av varandra. Elever inom en skola tenderade att vara mer lika än elever i andra skolor. Förhållandet att elever är grupperade i olika skolor (kluster) kan beskrivas som en sorts hierarki, där den första nivån är eleverna och den andra nivån är skolorna.

Observationer i form av kluster är inte ovanligt inom interventionsforskning och det beror ofta på att klustren utgörs av naturliga grupper eller på urvalsförfarandet. Faktum är att många typer av data, inklusive observationsdata inom hälsoforskning, har en *naturlig hierarki*. En grupp av patienter som går till en särskild läkare kan exempelvis skilja sig från en annan patientgrupp som går till en annan läkare. Skillnaderna kan ha sin grund i det område läkaren är verksam inom, i läkarens personliga egenskaper etcetera.

Urvalsförfarande påverkar också i vilken utsträckning data är i form av kluster; till exempel när vi har sjukhus eller mottagningar som urvalsenhet och sedan samlar in data för alla eller en undergrupp av patienter inom dessa grupper.

Experimentella forskningsdesigner kan också ge upphov till hierarkier i data. I randomiserade studier kan varje individ randomiseras till en interventionsgrupp. Interventionen ges direkt till individen och observationer kan göras av varje individ. Men ibland är det inte praktiskt möjligt med individuell randomisering utan grupper av individer måste utgöra randomiseringsenhet (klusterrandomisering). Att låta alla individer i samma grupp få samma intervention är också ett sätt att undvika ”smittoeffekter” mellan individer i interventions- och kontrollgrupper. I sådana situationer är data indelade i kluster genom studiens design. Ett exempel på en studie som använde klusterrandomisering är den svenska *Two-County trial* av bröstcancerscreening, som startade 1977 i Sverige. På grund av logistikproblem med mobila mammografienheter var individuell randomisering inte praktiskt möjlig. De två landstingen i Kopparbergs och Östergötlands län delades därför in i geografiska områden som

randomiserades till antingen screening eller kontrollgrupp (Tabar, Fagerberg, Gad, Baldetorp, Holmberg & Gröntoft m.fl., 1985). Alla kvinnor i interventionsområdena mellan 40 och 74 år inbjöds att genomgå mammografiscreening. Kvinnorna jämfördes sedan med kvinnor i kontrollområdena. Men det visade sig att kvinnor i en ort var mer lika än ett slumpmässigt urval av kvinnor från de två landstingen. Det fanns två källor till variation i studien; en mellan kvinnor inom ett område och en mellan områden. I den här studien fick de individuella kvinnorna ta del av interventionen mammografi. I andra fall riktas interventionen till en hel grupp snarare än till individer och klustereffekten kan då bli mycket större. I utvärderingar av hälsoutbildningsprogram i skolan är exempelvis ofta skolklasser randomiseringsenhet och programmen implementeras ofta av lärare i hela klassen (Murray, Varnell & Blitstein, 2004). Klusterrandomiserade studier är således vanliga i forskning inom folkhälsa och klinisk-praktisk verksamhet.

I klusterdata kan observationernas varians delas in i två komponenter: observationernas varians inom kluster (inomgruppsvarians) och variansen av klustrens medelvärden mellan kluster (mellangruppsvarians). Utfallets varians (den totala variansen) är summan av variansen inom kluster och variansen mellan kluster.

Observationernas beroende inom kluster kan uppskattas med intraklasskorrelationskoefficienten (ICC):

$$\text{ICC} = \frac{\text{mellangruppsvarians}}{\text{mellangruppsvarians} + \text{inomgruppsvarians}}$$

Ju mindre variansen är inom grupper, desto större blir ICC. Intraklasskorrelationskoefficienten kallas även ”designeffekt” (eng. *design effect*) och tolkas som proportionen av den totala kvarstående variansen (eng. *total residual variation*) som beror på skillnader mellan kluster.

Korrelerade observationer har två relaterade konsekvenser. Den första är att sampelstorleken påverkas av klusterdata. Eftersom individer inom ett kluster liknar varandra ger de mindre information

än om samma antal individer var ett urval ur den allmänna populationen. I praktiken förloras en del statistisk power på grund av randomisering av kluster i stället för individer – något som man bör ta hänsyn till när man beräknar storleken på sitt urval. För att beräkna hur många individer som behövs i en flernivåstudie måste man generellt sett först beräkna hur många som krävs i en standardstudie och sedan multiplicera det antalet med en *korrektionsfaktor* eller *inflationsfaktor* IF= [1+(*n*-1)ϱ], där *n* är genomsnittet av antal observationer per kluster och ϱ är intraklasskorrelationskoefficienten (Murray, 1998).

Den andra konsekvensen är att klusterdata påverkar studiens inferens, det vill säga möjligheten att dra slutsatser. Den hierarkiska strukturen gör att det allmänna antagandet om oberoende mätvärden som gäller för alla statistiska standardanalyser inte längre håller. Deltagarna i studien kan inte längre ses som oberoende individer. Vi måste därför ta hänsyn till variationen mellan kluster. Att bortse från hierarkier i data leder till underskattade standardfel i regressionsparametrarna och därmed till felaktigt snäva konfidensintervall, vilket ökar risken att förkasta nollhypoteser (att det inte finns någon interventionseffekt). Ju större och färre kluster, desto större blir underskattningen av standardfelen. Att bortse från sådant beroende i standardanalyser av linjär regression kan leda till felaktig statistisk inferens, och specifikt till felaktiga positiva resultat (Altman & Bland, 1997).

Data i fler nivåer måste med andra ord analyseras korrekt. Det enklaste sättet är att göra analysen på klusternivå genom att summera statistiken för varje kluster och sedan analysera summavärdena. Men en aggregerad analys på klusternivå är inte den bästa lösningen. Samband på aggregerad nivå kan över huvud taget inte användas som en uppskattning av samband på individnivå och generellt sett kan tolkningar av samband bli problematiska (Woodhouse & Goldstein, 1988). Dessutom är precisionen i skattningarna från en aggregerad analys lägre än skattningar från en flernivåanalys som baseras på individuella data (Goldstein, 2003).

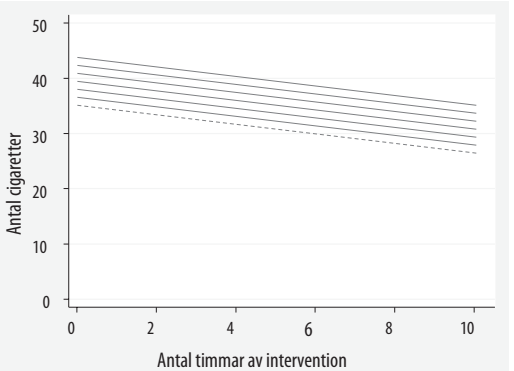
Ett alternativ är att anamma teknikerna för flernivåanalys. Det är statistiska tekniker som tar hänsyn till den hierarkiska strukturen i data. Flernivåanalys är en utvidgning av ”standardtekniker” för regression; flernivåanalys med en kontinuerlig utfallsvariabel är en utvidgning av linjär regressionsanalys och flernivåanalys medan en dikotom utfallsvariabel är en utvidgning av logistisk regressionsanalys.

Alla antaganden som gäller för ”standardregression” gäller följaktligen även för flernivåanalys. Det enda undantaget är antagandet om oberoende observationer, eftersom man gör flernivåanalyser just för att lösa problemet med korrelerade observationer. Flernivåanalyser har i stället ytterligare ett antagande som förklaras närmare i nästa avsnitt, att regressionslinjernas slumpmässiga intercept och lutning måste vara normalfördelade.

Linjär flernivåregression

Innan vi förklarar de grundläggande principerna i flernivåanalys visar vi tillämpningen av flernivåanalys på en kontinuerlig utfallsvariabel. I vår studie som utvärderar effekten av ett preventivt hälsoutbildningsprogram mot rökning vill vi nu ta med klustren av ungdomar (nivå 1) i skolor (nivå 2), det vill säga korrigera den linjära regressionsanalysen för enheten skola. För att undersöka sambandet mellan antalet utbildningstimmar och antalet cigaretter som ungdomarna rökte under månaden efter interventionen kan vi använda modell 2 och lägga till skola som korrigeringsvariabel. Nu är skola en kategorisk variabel och eftersom det är 20 skolor i studien behöver vi skapa 19 dummyvariabler som ska skattas i modellen. Att korrigera för skola innebär att 19 ytterligare koefficienter för intercepten måste beräknas i analysen (figur 14:3).

Det är slöseri med både statistisk power och effektivitet. Om det inte finns något särskilt intresse av att veta genomsnittligt antal rökta cigaretter i varje skola kan vi i stället för att skatta alla intercept endast skatta interceptens varians. Den variansen kallas varians



Figur 14:3. Regressionslinjer för sambandet mellan antalet rökta cigaretter under månaden efter interventionen och antal utbildningstimmar, korrigerade för skola.

på nivå 2 för att skolor är enheter på nivå 2. Med skattning av interceptens varians menas också att ”låta intercepten variera slumpmässigt”, alltså slumpvisa intercept. Det är anledningen till att flernivåanalys även kallas analys av slumpeffekter. Skolorna betraktas som ett slumpmässigt urval från en stor underliggande population av skolor. Det är inte de individuella skolorna i samplet som är i fokus i flernivåanalyser utan det centrala är att skatta variationsmönstret i den underliggande populationen av skolor. Flernivåmodellen blir då:

$$\text{Antal cigaretter}_{ij} = \beta_{\circ j} + \beta_1 \text{ antal utbildningstimmar}_{ij} + e_{ij} \qquad \text{(modell 8)}$$

där *i* är elever och *j* är skolor. Om *i* eller *j* saknas betyder det att parametern är konstant bland elever och skolor. För den här modellen gäller särskilt att när *j* läggs till interceptet betyder det att interceptet tillåts variera slumpmässigt på skolnivån.

Vi kan då uttrycka $\beta_{\circ j}$ som $\beta_\circ + u_{\circ j}$. $\beta_\circ$ är den underliggande populationens genomsnittliga intercept (en konstant) och $u_{\circ j}$ är skillnaden mellan den j:e skolans intercept och $\beta_\circ$. Den slumpmässiga kvantiteten $u_{\circ j}$ antas vara normalfördelad. Denna nya komponents varians, betecknad σ_u^2, återspeglar variansen av de olika skolornas intercept och representerar därmed ”effekten av enheten skola”.

Modellen ger de skattade värdena $\beta_\circ = 40$ (0,24), $\beta_1 = -1{,}6$ (0,07),

σ_e^2= 11,0 och σ_u^2 = 1,9. Om vi går tillbaka till resultaten för modell 2 (”standardmodellen” utan slumpvisa intercept) kan vi se att felvariansen i modell 8 är mindre. Det som har skett är att felvariansen i modell 2 (σ_e^2= 12,7) har i modell 8 fördelats på två komponenter motsvarande de två nivåerna i hierarkin: variansen mellan skolor, det vill säga variansen för komponenten på nivå 2 (σ_u^2 = 1,9), och variansen mellan ungdomar i en viss skola, det vill säga variansen för komponenten på nivå 1 (σ_e^2 = 11,0). En del av felvariansen i standardmodellen förklaras nu i modell 8 genom adderingen av slumpvisa intercept.

Likelihood ratio-test (LR-test) kan användas för att avgöra om det är nödvändigt att ha olika intercept för olika skolor. LR-test används för att jämföra hur bra två modeller appliceras till de befintliga data. En av modellerna (kallad den reducerade modellen) är ett speciellt fall av den andra, som kallas den kompletta modellen (eng. *full model*). LR-test beskriver hur pass sannolika de befintliga data är under den reducerade respektive under den kompletta modellen. Kvoten mellan dess två LR (närmare bestämt deras logaritm) kan användas för att bestämma om man ska välja bort den reducerade modellen och i stället anta den kompletta modellen. Testen görs genom att beräkna två gånger skillnaden mellan de två maximerade log-likelihood och jämföra den med en chi²-fördelning med antal frihetsgrader lika med skillnaden i antal parametrar mellan den kompletta och den reducerade modellen. I exemplet prövas skillnaden mellan standardmodellens och flernivåmodellens -2 log-sannolikheter L= 2*((-1839.6077)- (-1848.3734)) =17.5314 som följer en chi²-fördelning med en frihetsgrad, eftersom bara interceptens varians har lagts till i flernivåmodellen jämfört med standardmodellen. Skillnaden är starkt signifikant, vilket betyder att slumpvarians för intercepten ska finnas med i den linjära regressionsmodellen.

Intraklasskorrelationen kan beräknas genom att dela *mellan*skolsvariansen med summan av *mellan*skolsvariansen och *inom*skolsvariansen. I det här exemplet är ICC: 1,9/(1,9 + 11,0) = 0,15. Skola förklarar med andra ord 15 procent av den totala variationen av antalet rökta

cigaretter. I de flesta tvärsnittsstudier är ICC inte högre än 0,2.

Interventionseffekten är precis densamma i flernivåmodellen (d.v.s. -1,6), men standardfelet ökar från 0,04 till 0,07. Ett konfidensintervall på 95 % för interventionseffekten beräknas som -1,6 ± 1,96 * 0,07 med gränserna -0,77 till -1,50, som fortfarande är signifikant, men vidare än i modell 2. Det beror på att i standardanalysen är antagandet att observationerna av varje ungdom är oberoende och tillför 100 procent ny information för varje observation. Flernivåanalysen däremot har korrigerats för skola, vilket innebär att information från en elev i samma skola tillför mindre än 100 procent ny information. Ju högre korrelationskoefficient inom en skola, desto lägre grad av ny information från en elev i den skolan och desto högre standardfel i flernivåanalysen jämfört med standardanalysen.

Flernivåanalyser kan utökas och anpassas till andra situationer. Utöver slumpintercept kan man exempelvis låta regressionslinjernas lutning variera med slumpen. Klusterdata kan också vara grupperade i fler än två nivåer (t.ex. skolor i olika rektorsområden) som då utgör ytterligare en hierarkisk nivå i data. Kluster på en högre nivå hanteras på samma sätt som kluster av elever i skolor.

Logistisk flernivåregression

De generella principerna för logistisk flernivåanalys är desamma som dem som beskrivits för linjär flernivåanalys. Som illustration av logistisk flernivåanalys kan vi återvända till exemplet med ”standardmodellen” för logistisk regression. För att undersöka om en preventiv intervention mot bruk av alkohol har någon effekt på förekomst av berusning under månaden efter interventionen kan vi använda modell 7, men den här gången ta med att observationerna är grupperade i kluster av skolor.

$$\text{logit}\,(p_{ij}) = \beta_{0j} + \beta_1\,\text{intervention}_{ij} \qquad \text{(modell 9)}$$

Med *j* som har lagts till interceptet menas att ett slumpintercept är

med i modellen på nivån för skola. Vi kan då uttrycka β_{oj} som β_{o} + u_{oj}. Modellen ger de skattade värdena β_{o} = -1,7 (0,15), β_{1} = -0,6 (0,22) och σ_{u}^{2} = 0,15. Sannolikhetskvoten som jämför standardmodellen av logistisk regression med flernivåmodellen som låter interceptet variera slumpvis mellan skolor är L=2*(l1-l0)= 2* ((-563.33855) – (-568.49508)) = 10.31306 med en frihetsgrad (för den adderade parametern σ_{u}^{2}). Eftersom skillnaden är signifikant drar vi slutsatsen att det finns signifikant variation mellan skolor och att det är nödvändigt att korrigera för klusterenheten skola i analysen.

Oddsen för att rapportera berusning för interventionsgruppens ungdomar är $e^{\beta 1}$ = 0,5 gånger oddsen för ungdomarna i kontrollgruppen, vilket är precis samma oddskvot som i modell 7. Standardfelet för interventionseffekten är som väntat större i flernivåanalysen än i analysen som ignorerar kluster (0,22 mot 0,16). Följaktligen har konfidensintervallet på 95 procent gränserna 0,35 till 0,87, vilket betyder att det är vidare än i modell 7. Men det är fortfarande signifikant skilt från 1. Oddskvotens signifikans fås från p-värdena i Walds test: $(-0,6/0,22)^{2}$ = 6,6, som är signifikant.

Det kan ifrågasättas om ICC ska skattas i logistisk flernivåanalys, eftersom en korrelationskoefficient för en dikotom variabel inte kan tolkas.

Sammanfattning

Vilken typ av utfall har du?

- Om utfallet är naturligt dikotomt: använd logistisk regression.
- Om utfallet är kontinuerligt: kontrollera normalitet i fördelningen.
- Om utfallet är kontinuerligt och normalfördelat: använd linjär regression.
- Om utfallet är kontinuerligt men inte normalfördelat: dikotomisera och använd logistisk regression.

Hur mäts interventionen?

- Om interventionen mäts på en kontinuerlig skala: använd regressionsanalys.
- Om interventionen är binär: använd antingen regression eller ANOVA/ANCOVA.

Mäts utfallet på individnivå men varje individ ingår i ett kluster?

- Använd en standardmodell av regression för att uppskatta effekten av interventionen med endast interventionen som oberoende variabel. Gör sedan om analysen med samma modell men med korrigering för kluster, det vill säga tillåt interceptet att variera slumpmässigt på klusternivån.
- Om sannolikhetskvoten som jämför de två modellerna är signifikant är det nödvändigt att använda flernivåanalys. Om inte, kan standardmodellen av regression användas.

Fördjupningslitteratur

Goldstein, H. (2003). *Multilevel statistical models*. London: Arnold.

Hardy, M. A. (1993). *Regression with dummy variables*. Newbury Park, Calif: Sage.

Kleinbaum, D. G. (2007). *Applied regression analysis and other multivariable methods.* Australia: Brooks/Cole.

Twisk, J,W.R. (2006). *Applied multilevel analysis*. Cambridge: Cambridge University Press.

Referenser

Altman, D. G. & Bland, J. M. (1997). Statistics notes. Units of analysis. *BMJ, 314,* 1874.

Goldstein, H. (1987). *Multilevel models in educational and social research.* London: Griffin.

Goldstein, H. (2003). *Multilevel statistical models.* London: Arnold.

Hardy, M. A. (1993). *Regression with dummy variables*. Newbury Park, Calif: Sage.

Jacobs, N., De Bourdeaudhuij, I., Thijs, H., Dendale, P. & Claes, N. (2011). Effect of a cardiovascular prevention program on health behavior and BMI in highly educated adults: A randomized controlled trial. *Patient Educ Couns, 85*, 122–6.

Murray, D. M., Varnell, S. P. & Blitstein, J. L. (2004). Design and analysis of group-randomized trials: a review of recent methodological developments. *American Journal of Public Health, 94*, 423–432.

Murray, M. (1998). *Design and Analysis of Group-Randomized Trials.* New York: Oxford University Press Inc.

Stanley, K. (2007). Design of randomized controlled trials. *Circulation, 115,* 1164–1169.

Tabar, L., Fagerberg, C. J., Gad, A., Baldetorp, L., Holmberg, L. H., Gröntoft,

O., m.fl. (1985). Reduction in mortality from breast cancer after mass screening with mammography. Randomised trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. *Lancet, 1*, 829–832.
Woodhouse, G. & Goldstein, H. (1988). Educational Performance Indicators and Lea League Tables. *Oxford Review of Education, 14*, 301–320.

Kapitel 15 | Lars-Göran Öst & Rikard Wicksell

Moderatorer, mediatorer och verkningsmekanismer

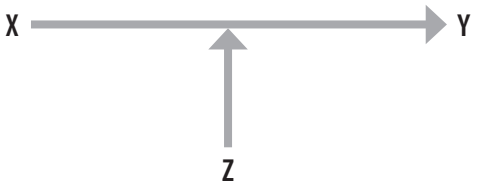
Definitioner av begreppen

Baron och Kenny (1986) publicerade en sedermera klassisk artikel om moderator- och mediatorvariabler ur begreppslig, strategisk och statistisk synvinkel och definierade begreppen. Denna artikel hade citerats 14 465 gånger den 1 december 2010.

Moderator

En moderator är enligt Baron och Kenny (1986) en kvalitativ (t.ex. kön, etnisk tillhörighet, socialklass) eller kvantitativ (t.ex. symtomens svårighetsgrad före behandlingen) variabel som påverkar riktningen av eller styrkan på relationen mellan en oberoende (eller prediktor-) variabel (behandling) och en beroende (eller kriterie-) variabel (utfall). En grundläggande moderatoreffekt kan beskrivas som en interaktion mellan en oberoende variabel och en faktor (moderatorn) som specificerar (klargör) de betingelser under vilka effekten fungerar.

Holmbeck (1997) säger att en moderatorvariabel är en som påverkar relationen mellan två variabler så att hur behandlingen på-



Figur 15:1. Modell som illustrerar moderatoreffekt (X = oberoende variabel, Y = beroendevariabel, Z = moderatorvariabel).

verkar kriterievariabeln varierar beroende på nivån hos eller värdet på moderatorn.

Frågan ses ur ett behandlingsperspektiv av Kraemer, Wilson, Fairburn och Agras (2002). De beskriver att behandlingsmoderatorer specificerar för vem eller under vilka betingelser behandlingen fungerar och att moderatorer kan identifiera subgrupper som möjligen har olika orsaksmekanismer eller förlopp på störningen.

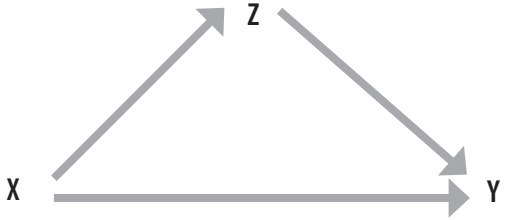
Gemensamt för moderatorer är att det är faktorer som finns hos individen eller miljön före en eventuell intervention (t.ex. att någon fått terapi i en behandlingsstudie) och att den också bör mätas före interventionen. Moderatoreffekten illustreras i figur 15:1.

Mediator

Baron och Kenny (1986) beskriver en mediator som den genererande mekanism genom vilken den oberoende variabeln kan påverka beroendevariabeln av intresse och att mediering sker bäst när det finns ett starkt samband mellan prediktorn och kriterievariabeln.

Holmbeck (1997) förtydligar genom att säga att den oberoende variabeln påverkar mediatorn som i sin tur påverkar den beroende (utfalls-)variabeln. Han betonar också att det som en viktig förutsättning ska finnas ett signifikant samband mellan oberoende och beroende variabel innan man testar för en mediatoreffekt.

Kazdin (2007) har ett psykoterapiperspektiv när han beskriver mediator som en inskjuten variabel som statistiskt kan förklara relationen mellan den oberoende och den beroende variabeln. Något som medierar förändring behöver dock inte nödvändigtvis förklara



Figur 15:2. Modell som illustrerar moderatoreffekt (X = oberoende variabel, Y = beroendevariabel, Z = mediatorvariabel).

processerna för hur förändringen skedde. Kraemer, Wilson, Fairburn och Agras (2002) säger också att alla mediatorer inte är verkningsmekanismer men alla mekanismer är mediatorer. Mediatoreffekten illustreras i figur 15:2.

Verkningsmekanism

Kazdin (2007) påpekar att verkningsmekanism är grunden för behandlingseffekten, det vill säga de processer eller händelser som är ansvariga för förändringen. Man kan också beskriva det som skälen till varför en förändring inträffade eller hur förändringen skedde. Till skillnad från mediatorer som illustrerar ett statistiskt samband med oberoende respektive beroende variabel kräver verkningsmekanismer att man kan uttala sig om kausalitet, det vill säga en faktor som orsakar att behandlingens effekt kommer till stånd.

Varför är dessa variabler intressanta att undersöka?

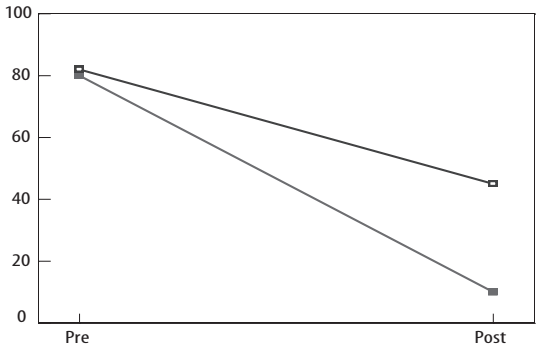
Kraemer med flera (2002) anser att kunskap om moderatorer kan ge värdefull information för att vägleda framtida omstrukturering av diagnostisk klassifikation (t.ex. DSM-5) och att fatta beslut om behandling. Det senare illustreras av olika försök att matcha specifika behandlingsmetoder till olika patientkarakteristiska, till exempel Project MATCH (1998) vid behandling av alkoholberoende. Inom fobiforskningen kan nämnas Östs, Jerremalms och Johans-

sons (1981; 1982; 1984) studier där olika responsmönster (fysiologiskt, beteendemässigt respektive kognitivt) matchades med teoretiskt sett passande respektive icke-passande behandlingsmetoder.

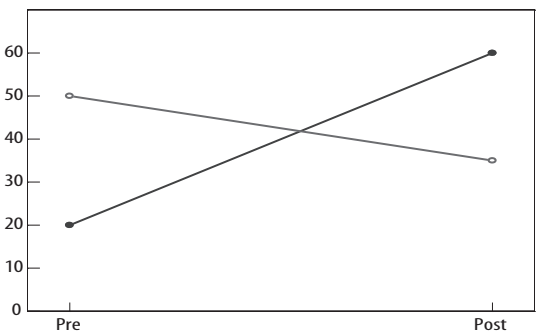
Den som kanske tydligast beskrivit varför det är viktigt att forska om mediatorer och verkningsmekanismer är Kazdin (2007). Han tar upp flera skäl:

1. Det finns i dag ett mycket stort antal mer eller mindre olika behandlingsmetoder och förståelse av förändringsmekanismer kan bringa ordning och sparsamhet i denna ”röra”.
2. Utveckling av terapiers mekanismer kommer att klargöra sambanden mellan vad som görs (behandlingen) och olika utfall.
3. Genom att förstå processerna som ansvarar för den terapeutiska förändringen bör man på ett bättre sätt kunna optimera behandlingen. Om vi känner till hur förändringen sker kanske vi kan rikta bättre, starkare, annorlunda eller flera strategier som sätter igång de kritiska förändringsprocesserna.
4. För att optimera spridningen av nya behandlingsmetoder från forskningsmiljöerna till klinisk rutinvård är det önskvärt att veta vad som behöver göras för att behandlingen ska fungera.
5. Förståelse av hur behandlingen fungerar kan också hjälpa till att identifiera moderatorer. Ett ytterligare skäl som Kazdin inte beskriver är att utbildningen av blivande psykoterapeuter kan optimeras om vi känner en behandlingsmetods verkningsmekanism. Vi kan då fördela tid till det som är viktigt för utfallet och inte ägna tid åt sådant som antagligen inte är betydelsefullt för utfallet.

Kraemer med flera (2002) tar också upp det kanske oftast anförda skälet till varför det är viktigt att forska om verkningsmekanismer. De säger att förståelse av de mekanismer som behandlingen verkar genom sannolikt kommer att underlätta utvecklingen av innovativa behandlingar, eller modifieringar av gamla metoder, som kommer att ge större eller samma effektstorlekar (ES) som nu men till lägre kostnad. Det måste dock påpekas att detta fortfarande är en from förhoppning då vi ännu inte känner verkningsmekanismen för nå-



Figur 15:3. Exempel på ordinal interaktionseffekt.



Figur 15:4. Exempel på disordinal interaktionseffekt.

gon behandlingsmetod och således inte har ”facit” om sådan kunskap verkligen leder till förbättrade behandlingar.

Moderatoranalys

Fairchild och McQuillin (2010) beskriver att moderatoreffekter kan vara av två typer: ordinala och disordinala interaktioner. När man redovisar data i figurer så illustreras ordinala interaktioner av linjer som inte korsar varandra (figur 15:3) medan disordinala interaktioner kännetecknas av linjer som korsar varandra (figur 15:4).

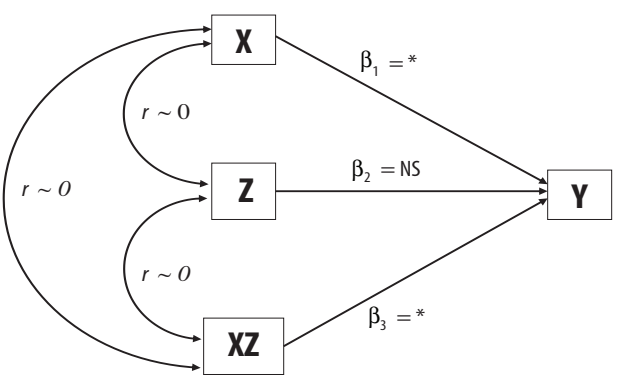
Den grundläggande moderatormodellen

Följande ekvation för multipel regression estimerar den grundläggande moderatormodellen:

$$(1) \qquad Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ$$

Y är det predicerade värdet av patientens prestation, X representerar gruppfördelningen (behandling jämfört med kontroll), Z representerar patientens värde på moderatorvariabeln och XZ är interaktionen mellan den oberoende variablen och moderatorvariabeln (termen bildas genom att multiplicera X och Z).

Regressionskoefficienterna i ekvationen kvantifierar effekten av en variabel medan effekten av övriga variabler kontrolleras i modellen. Den grundläggande moderatormodellen illustreras i figur 15:5. Av denna framgår att korrelationerna mellan prediktorvariablerna (X, Z, och XZ) är approximativt 0 medan den oberoende variabeln (X) korrelerar signifikant med beroendevariabeln (Y) när effekten av moderatorvariabeln (Z) och interaktionen (XZ) kontrolleras för i regressionsanalysen. Moderatorvariabeln (Z) predicerar däremot inte utfallsvariabeln (Y) när effekten av den oberoende variabeln (X) och interaktionen (XZ) kontrolleras för. Slutligen framgår att interaktionen (XZ) mellan den oberoende variabeln (X) och moderatorvariabeln (Z) signifikant predicerar utfallet (Y) när lägre ordningens effekter kontrolleras i regressionsanalysen.



Figur 15:5. Den grundläggande moderatormodellen (X = oberoende variabel, Y = beroendevariabel, Z = moderatorvariabel, XZ interaktionen mellan oberoende och moderatorvariabel, β_1 = effekten av X på Y när Z och XZ kontrolleras, β_2 = effekten av Z på Y när X och XZ kontrolleras, β_3 = effekten av XZ på Y när X och Z kontrolleras).

Testning av moderatoreffekt

Ett *t*-test av regressionskoefficienten som hör till XZ-interaktionen är ett sätt att bestämma om det finns en statistisk moderation. Om β_3 är signifikant finns en signifikant moderatoreffekt.

Man kan också använda ett R^2-förändringstest (betecknas som R^2 Δ) för att evaluera moderation (Aiken & West, 1991). R^2 Δ-testet är fördelat som *F*-värdet och kvantifierar den varians som förklaras av XZ-interaktionen, utöver den varians som förklaras i modellen utan att interaktionstermen är inkluderad. Detta är identiskt med resultatet av *t*-testet för β_3 så att kvadratroten ur *F*-värdet för R^2 Δ-testet är lika med *t*-värdet och *p*-värdena är identiska.

Tolkning av moderatoreffekter

Relationen mellan X och Y måste analyseras vid olika värden på moderatorvariabeln Z för att förstå sambandet. Att plotta upp interaktionen i en figur hjälper till vid tolkningen av interaktionseffekten och innebär ett sätt att undersöka hur relationen mellan X och Y förändras över olika nivåer på moderatorvariabeln. När man plottar upp interaktioner måste specifika värden på moderatorvariabeln väljas för vilka sambandet mellan X och Y ritas upp. Om det är en kategorivariabel ska värdena på Z motsvara de grupper som man vill rita upp. När det är en kontinuerlig moderatorvariabel bör teoretiska kriterier ge meningsfulla värden på Z och dessa bör användas. Finns det inga sådana värden rekommenderar Aiken och West (1991) att man plottar upp enkla kurvor för -1SD, M och +1SD av Z.

Regressionslinjen som motsvarar prediktionen av Y från X vid ett enskilt värde av Z kallas för enkla lutningar. Aiken och West (1991) beskriver statistiska test för att pröva om regressionslinjen är signifikant skild från noll. De illustrerar också hur man kan testa om par av regressionslinjer skiljer sig signifikant från varandra.

Effektstorlek i moderatormodeller

Fairchild och McQuillin (2010) beskriver effektstorlek som den praktiska betydelsen av en effekt. Mått på effektstorlek är viktiga

när man tolkar analysresultat då de kan ge information om icke-signifikanta fynd och öka förståelsen för den praktiska användbarheten av statistiskt signifikanta effekter. Exempel på effektstorleksmått vid moderatoranalys är kvadrerad partiell korrelationskoefficient (partiell r^2), vilken är ekvivalent med $R^2\Delta$, som illustrerar den unika andel av variansen som förklaras av en prediktor till utfallsvariabeln (kriteriet) och som inte förklaras av andra prediktorer i modellen. Cohen (1988) definierade en liten kvadrerad partiell r^2 som 0.02–0.12, en måttlig som 0.13–0.25 och en stor som ≥0.26.

Power i moderatormodeller

Den statiska powern är ofta liten i moderatoranalyser. Fairchild och McQuillin (2010) beskriver att översikter av frågan inom socialvetenskap ofta finner att interaktionseffekten förklarar mellan 1 och 3 procent av variansen i beroendevariabeln. För att maximera powern att upptäcka moderatoreffekter rekommenderar de att man ska använda så stora sampel som möjligt, använda extremgrupper för att öka variansen i designen och välja mått som har hög reliabilitet.

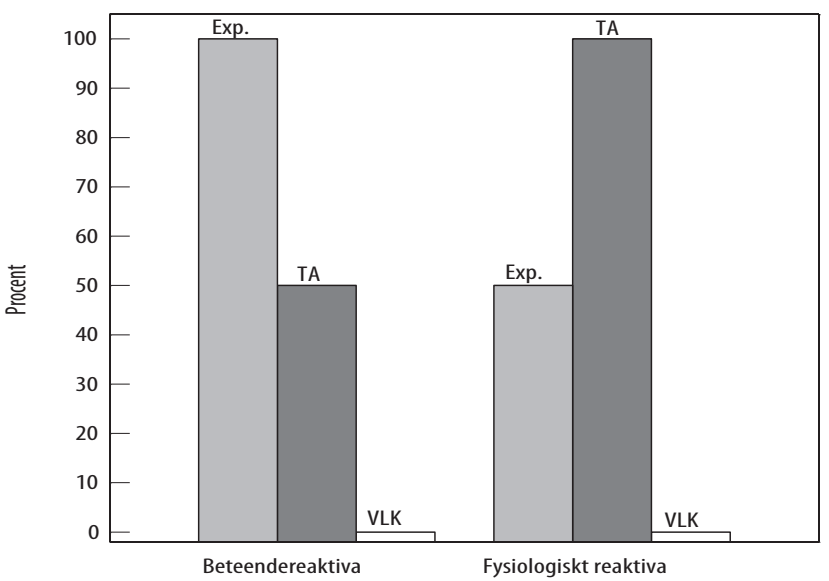
Antaganden för moderatormodellen

De flesta antaganden som ligger bakom moderatormodellen är de som vanlig regressionsanalys utgår från som korrekt modellspecifikation, mätningar utan mätfel och normalfördelade residualer. Ett av de viktigaste antagandena vid moderatoranalys är homogenitet hos felvariansen (homoscedasticitet). Detta innebär att residualvariansen som återstår efter att ha predicerat Y från X är konstant över alla värden på moderatorvariabeln, Z. Bartletts M är ett statistiskt test av om kravet på homogenitet hos felvariansen har uppfyllts eller inte. Skulle detta indikera heteroscedasticitet finns det icke-parametriska test för att testa moderation (Fairchild & McQuillin, 2010).

Exempel på moderatoranalys i behandlingsforskning

Det är sedan länge känt att en fobi manifesteras i tre olika kompo-

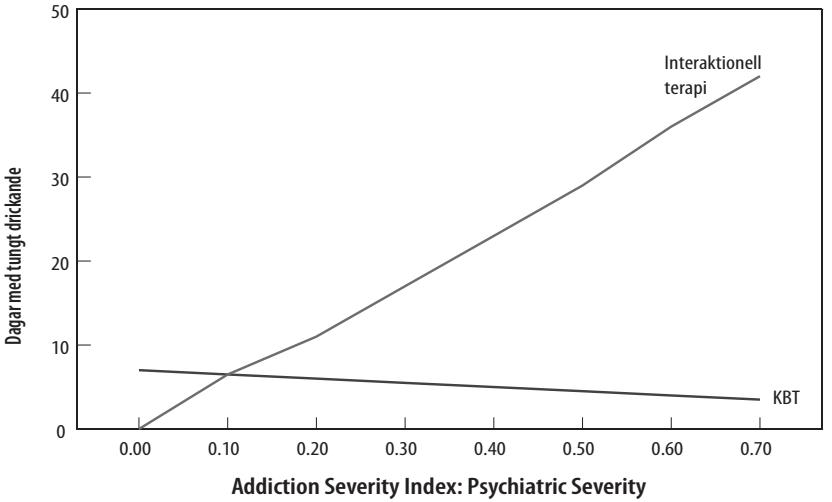
nenter: beteende (undvikande eller flykt), fysiologiskt påslag (t.ex. pulsökning) och negativa kognitioner (t.ex. katastroftankar). I en studie av klaustrofobi undersökte Öst, Johansson och Jerremalm (1982) effekten av att matcha behandlingsmetod till patientens mest framträdande reaktionsmönster. Som en del av förmätningen fick patienterna gå in i ett litet rum (75 × 120 cm) utan fönster och låsa dörren till rummet. Under testningen mättes undvikandebeteende, hjärtfrekvens och subjektiv upplevelse av ångest. De patienter som hade starkt undvikandebeteende men liten ökning av pulsen kategoriserades som beteendereaktiva medan de som fick en stark pulsökning men hade lite undvikandebeteende bedömdes som fysiologiskt reaktiva. Inom vardera av dessa responsgrupper randomiserades patienterna till åtta sessioners behandling med exponering



Figur 15:6. Andelen kliniskt signifikanta förbättrade efter exponering (Exp.), tillämpad avslappning (TA) och väntelista (VLK) bland beteende- respektive fysiologiskt reaktiva klaustrofobiska patienter (Öst m.fl., 1982).

in-vivo (E) eller tillämpad avslappning (TA) respektive ingen behandling (väntelistekontroll, VLK). Hypotesen var att för de beteendereaktiva skulle E ge signifikant bättre resultat än TA och bägge metoderna skulle vara bättre än VLK. För de fysiologiskt reaktiva däremot skulle TA vara signifikant bättre än E och bägge bättre än VLK. Kliniskt signifikant förbättring utgjordes av stora förändringar på två mått samtidigt och resultaten illustreras i figur 15:6. För både beteendereaktiva och fysiologiskt reaktiva patienter ledde den behandlingsmetod som matchade patienternas responsmönster (E respektive TA) till att 100 procent blev förbättrade medan den icke-matchade metoden (TA respektive E) endast ledde till 50 procent och VLK till ingen förändring. Denna studie illustrerar tydligt hur ett kännetecken på klienterna (responsmönster) modererar behandlingsresultaten.

Den modererande effekten av en kontinuerlig variabel, psykiatrisk svårighetsgrad före behandlingen (mätt med Addiction Severity Index), undersöktes av Kadden, Cooney, Getter och Litt (1989) hos ett sampel av 96 alkoholberoende patienter. Dessa randomiserades till två typer av eftervård som gavs i grupp; KBT (coping skills program) eller interaktionell terapi (Yaloms form av gruppterapi). Man undersökte sedan drickande hos patienterna under en uppföljningsperiod på sex månader. Figur 15:7 illustrerar resultatet analyserat med hierarkisk linjär regressionsanalys. För patienterna som fått KBT hade den psykiatriska svårighetsgraden före behandlingen ingen effekt på drickandet, men för dem som fick interaktionell terapi blev effekten sämre (mer drickande) ju högre psykiatrisk svårighetsgrad patienterna hade. Denna studie visar alltså en modererande effekt för en av de två behandlingsmetoderna. När man i denna studie analyserade det dikotoma måttet drickande–inget drickande med logistisk analys blev resultatet diametralt motsatt för behandlingsgrupperna. För de patienter som fått interaktionell terapi så ökade sannolikheten för återfall (drack under uppföljningsperioden) ju högre psykiatrisk svårighetsgrad man hade före behandlingen. För de KBT-behandlade patienterna var mönstret det motsatta: sanno-



Figur 15:7. Andelen dagar med tungt drickande efter interaktionell terapi respektive KBT modererat av graden av psykiatrisk svårighet före behandlingen (Kadden m.fl., 1989).

likheten för återfall *minskade* när den psykiatriska svårighetsgraden före behandlingen ökade.

Mediatoranalys

Vad är en mediator?

Till skillnad från en moderator är mediatorn (M) en del av en sekvens och utgör en länk i den kedja som sammanbinder X och Y genom att X predicerar M som i sin tur predicerar Y (Fairchild & McQuillin, 2010). Sambandet mellan X och Y kan beskrivas som en generativ process (X genererar en förändring i Y), i vilken M representerar en betydelsefull faktor (Baron & Kenny, 1986) som bidrar till att förklara detta samband. Analyser av mediatorer är således en kartläggning av den process där X påverkar Y (MacKinnon, Fairchild & Fritz, 2007). Denna typ av forskning kallas av denna anledning därför ofta för just processforskning.

Woodworths S-O-R-modell

Ett av de första och kanske tydligaste exemplen på hur en tredje variabel kan inkluderas i en teoretisk modell i formen av en mediator är Woodworths beskrivning av sambandet mellan stimulus-organism-respons (MacKinnon, 2008). Woodworth utvecklade en modell i vilken han beskrev att stimulus (X) initierar aktiviteter i organismen (M) och att dessa förändringar bidrar till att generera effekter på beteendet (Y). Ett visuellt stimulus, X (lejon), leder till en mental aktivitet, M (tolkning om fara, baserat på tidigare erfarenheter och andras berättelser), som i sin tur ligger till grund för beslutet om hur man klokast agerar, Y (springer därifrån). Att beskriva Y som ett resultat av X är inte felaktigt, men inkluderingen av M ger en mer detaljerad bild av processen.

Indirekt och direkt effekt

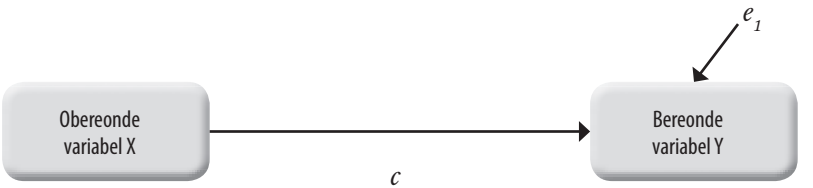
Variabeln X:s inverkan på Y kan vara av olika karaktär. I sammanhang där man inkluderar en tredje variabel i formen av en mediator (M) delar man vanligen upp effekten av X på Y i två delar: den direkta och den indirekta effekten. Tillsammans bildar dessa den totala effekten. Ett enkelt exempel illustrerar betydelsen av indirekt och direkt effekt. Eftersom statistiska analyser har påvisat ett starkt samband mellan en prediktor, smärta (X), och en utfallsvariabel, livskvalitet (Y), är man intresserad av att undersöka om detta samband går via en tredje variabel, oro för smärta (M). Analyserna visar på ett signifikant samband mellan dels (a) smärta och oro för smärta, dels (b) oro för smärta och livskvalitet (även vid kontroll för effekten av smärta). Detta talar för en indirekt effekt av smärta på livskvalitet som går via variabeln oro för smärta. Låt säga att kompletterande analyser i exemplet ovan visar att sambandet mellan smärta och livskvalitet kvarstår också när effekten av orosnivå är kontrollerad för. Detta indikerar att det även finns en direkt effekt av smärta på livskvalitet, det vill säga utöver den effekt som kan förklaras av mediatorn oro för smärta. Genom att på detta sätt inkludera en tredje variabel i analysen kan man förfina förståelsen

för sambandet mellan en prediktor och en utfallsvariabel genom att dela in påverkan av X på Y i den del som går via en potentiell mediator (indirekt effekt) och den del som inte förklaras av mediatorn (direkt effekt). Ofta förespråkas användandet av termen indirekt effekt i stället för medieringseffekt (MacKinnon, 2008; Preacher & Hayes, 2004).

Komplett och partiell mediering

En mediator fungerar alltså som en länk som förklarar en betydande del av relationen mellan X och Y. I det mest extrema fallet förklaras relationen mellan X och Y helt och hållet av den indirekta effekten, och man talar då om en komplett mediering som kännetecknas av att den direkta effekten (den del av sambandet som kvarstår när M kontrolleras för) blir 0. När detta förekommer visar analysen på en enskild och helt dominerande mediator, vilket i praktiken är ovanligt i exempelvis psykologisk och sociologisk forskning. Vanligare är att man i dessa sammanhang talar om partiell mediering och i stället försöker bedöma storleken, eller betydelsen, av den indirekta effekten – med andra ord, om införandet av variabeln M bidrar till en signifikant minskning av styrkan i sambandet mellan X och Y (MacKinnon, 2008).

Om den direkta effekten inte är 0, det vill säga om medieringen är partiell, kan detta också tolkas som en indikation på att det finns fler medierande faktorer. Att beskriva M som en länk i kedjan betyder inte att X och Y antas bindas samman av en enda variabel. Sambandet mellan prediktorn och beroendevariabeln kan gå flera vägar, vilket alltså indikeras av att den direkta effekten av X på Y inte är 0. Studier av indirekta effekter bör således beakta flera möjliga mediatorer. Detta kan göras på många olika sätt. Dels kan den relativa betydelsen av olika mediatorer undersökas, dels kan mer komplexa modeller testas där två eller fler mediatorvariabler ingår. Hur flera mediatorer kan, och ibland bör, beaktas samtidigt diskuteras senare i kapitlet.

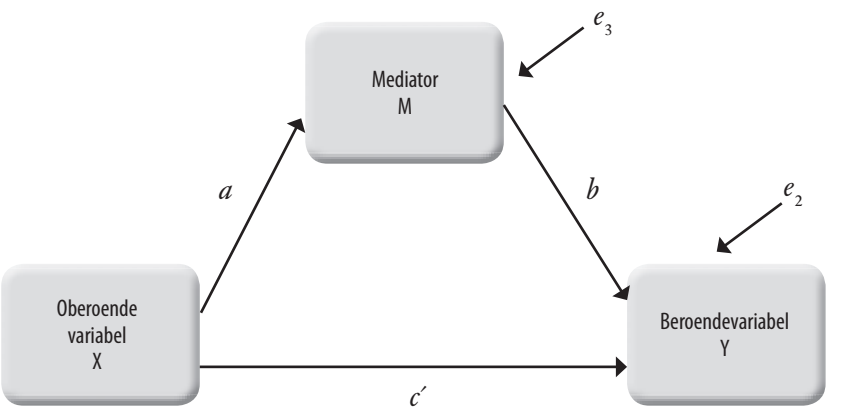


$$Y = i_1 + cX + e_1$$

Figur 15:8. Path diagram och ekvation. En modell som illustrerar relationen mellan en oberoende och en beroende variabel.

Den grundläggande mediatormodellen

Figur 15:8 representerar en enkel modell för hur en variabel (X) är relaterad till en annan (Y). I detta fall indikerar pilen att X utgör en prediktor för Y, och relationen mellan X och Y betecknas med symbolen *c* (MacKinnon, 2008), vilket kan sägas vara den totala effekten. Parametern *e* representerar den del av variationen (koefficienten) i Y som inte förklaras av X, den så kallade felvariansen (el-



$$Y = i_2 + c'X + b\,M_1 + e_2$$
$$M_1 = i_3 + a\,X + e_3$$

Figur 15:9. Path diagram och ekvationer. En modell som illustrerar sambandet mellan en oberoende, en medierande och en beroende variabel.

ler oförklarade variansen). I ekvationen förekommer även termen *i,* vilken står för interceptet.

I figur 15:9 har modellen utvidgats till att också innehålla en tredje variabel, en mediator.

Den oberoende variabeln utgör fortfarande en prediktor för Y, men som tidigare beskrivits går denna relation två vägar. Exempelvis påverkar X mediatorn (M) som i sin tur påverkar Y. Relationen mellan X och M betecknas *a*, och sambandet mellan M och Y (vid kontroll för X) symboliseras *b*. Relationen X-M-Y beskrivs som den indirekta effekten av X på Y, via mediatorn (M). Figuren illustrerar också den direkta effekten av X på Y, det vill säga den påverkan X har på Y när effekten av M är kontrollerad för. Den direkta effekten betecknas här *c´*. Som i figuren ovan utgör koefficienten e_2 den del av variationen i Y som inte förklaras av relationen till vare sig X eller M. På motsvarande sätt representerar koefficienten e_3 felvariansen i variabeln M, det vill säga den del av variansen i M som inte förklaras av X. Mer komplexa modeller kan konstrueras men detta utgör den enklaste, och huvudsakliga, mediatormodellen. Värt att notera är att även om modellen i figur 15:8 går att tillämpa i studier med olika design är den experimentella manipulationen av X en viktig förutsättning för att identifiera en mediator (Kazdin, 2007). Som framgår i figurerna ovan kan sambanden mellan X och Y respektive X, M och Y beskrivas med regressionsekvationer, och det är dessa som utgör grunden för de statistiska analyserna av medieringseffekter. I ekvationerna förekommer en ytterligare term, *i*, vilken representerar interceptet. Den termen behövs för att ekvationen ska vara komplett och kan exempelvis användas för att grafiskt åskådliggöra medieringseffekter. Dock är interceptet inte av någon egentlig betydelse vid beräkningen av medieringseffekter (MacKinnon, 2008). De tre regressionsanalyser som används för beräkning av medieringseffekter, och som förekommer i figurerna 15:8 och 15:9, är således följande:

$$(2) \qquad Y = i_1 + cX + e_1$$

$$(3) \qquad Y = i_2 + c'X + bM + e_2$$
$$(4) \qquad M = i_3 + aX + e_3$$

Som nämnts och som framgår av figur 15:8, representerar *c* effekten av X på Y, det vill säga den totala effekten på beroendevariabeln som uppstår när vi manipulerar den oberoende variabeln. Tidigare har man ansett att medieringseffekter förutsätter ett starkt samband mellan X och Y, med andra ord att koefficienten *c* är signifikant (Baron & Kenny, 1986). På senare tid har dock vissa forskare argumenterat starkt för värdet av att undersöka indirekta effekter av X på Y oavsett styrkan på sambandet, då mediatorer kan vara av stor betydelse för relationen mellan X och Y oavsett om denna är signifikant eller ej (MacKinnon, 2008; Preacher & Hayes, 2004).

Det är också av stor betydelse att hypoteser kring relevanta processvariabler är väl teoretiskt förankrade. Med andra ord bör medieringsanalyser innefatta väl formulerade hypoteser om vilka variabler som, utifrån teorin, bör förändras i den aktuella behandlingen och vilka variabler som inte förväntas förändras.

Olika sätt att mäta medieringseffekter

En konceptuell modell för mediering, inklusive ekvationer och diagram som tydliggör sambanden mellan X, M och Y, har nu presenterats. Analyser av mediatorer förutsätter också att man på något sätt kan kvantifiera styrkan i medieringseffekten. Utöver bedömningen om mediering kan tänkas förekomma, möjliggör detta också en relevant jämförelse mellan olika mediatorer eller mellan olika studier. Genom att mäta medieringseffekten ges även möjligheten att värdera denna i relation till den direkta eller den totala effekten av sambandet mellan X och Y. Här följer en övergripande beskrivning av några olika sätt att bedöma medieringseffekter.

Analys av de enskilda stegen i medieringsmodellen

Det finns ett flertal sätt att närma sig frågorna om det förekommer en medieringseffekt och, i så fall, hur stark denna är. Den metod

som tidigare var mest förespråkad och använd byggde på en serie av regressionsanalyser för att undersöka vart och ett av stegen i medieringsmodellen (Baron & Kenny, 1986). Först fastställs ett signifikant samband mellan X och Y, vilket i 15:8 betyder att koefficienten *c* måste vara signifikant. Därefter undersöks om X har en signifikant inverkan på M, det vill säga om den oberoende variabeln är av tillräckligt stor betydelse för mediatorn. Med andra ord behöver länken mellan X och M, *a*, vara signifikant. Vidare krävs att mediatorn har ett signifikant samband med beroendevariabeln vid kontroll för effekten av den oberoende variabeln. I figur 15:9 kan detta beskrivas som att koefficienten *b* måste vara signifikant. Detta är ett kritiskt steg då ett negativt resultat indikerar att den effekt som X har på M inte förs vidare till Y, utan att det snarare handlar om att X har en betydande påverkan på både M och Y, men var för sig. Slutligen måste den direkta effekten *c´* vara icke-signifikant. Detta betyder i praktiken att det samband mellan X och Y som påvisades vara signifikant, det vill säga *c* i figur 15:9, minskar påtagligt i styrka som ett resultat av att vi tar hänsyn till den effekt som går via mediatorn. I statistiska termer betyder detta att den bivariata korrelationen mellan X och Y är signifikant, men att den partiella korrelationen mellan X och Y där M kontrolleras för är icke-signifikant.

Som tidigare nämnts har kravet på en signifikant relation mellan X och Y fått en hel del kritik, framför allt av två skäl. Först och främst finns det ingen konceptuell eller teoretisk motsättning mellan å ena sidan ett icke-signifikant samband mellan X och Y, och å andra sidan förekomsten av en mediator. Det vill säga, även ett svagt samband mellan X och Y kan förklaras till större del av den indirekta än av den direkta effekten. Dessutom bidrar detta statistiska kriterium till att faktiska medieringseffekter aldrig upptäcks, då analysen ”stannar” vid att relationen X–Y inte är signifikant. I statistiska termer betyder detta att det föreligger en påtaglig risk för typ-I-fel (MacKinnon m.fl., 2007).

Skillnader i koefficienter

Mediering kan alltså beskrivas som den indirekta effekten av X på Y via M. Ett sätt att mäta medieringseffekten är att beräkna minskningen av styrkan i sambandet mellan X och Y när hänsyn tas till M. Genom att dra bort den direkta effekten *c´* från den totala effekten *c* så får man alltså fram ett mått på den indirekta effekten (*c - c´*). I händelse av att mediatorn svarar för en betydande del av relationen mellan X och Y kommer således *c´* att vara högt, vilket i sin tur medför att *c* reduceras betydligt. Det finns en tilltalande logik i detta tillvägagångssätt. Dock har metodologiska analyser visat på svårigheter att använda differensen mellan *c* och *c´* för att undersöka medieringseffekter, framför allt vid mer komplexa modeller som exempelvis innefattar flera mediatorer (Fairchild & McQuillin, 2010; MacKinnon, 2008).

Produkten av koefficienterna

Effekten av X på M samt effekten av M på Y vid kontroll för X kan också beskrivas som kombinationen av relationerna *a* och *b*. Med andra ord kan vi multiplicera *a* och *b* för att få fram styrkan i den indirekta effekten (*ab*). Produkten av koefficienterna *ab* beskriver således hur mycket en viss förändring i X förändrar Y genom effekter på M. På motsvarande sätt visar förändringen i *c - c´* hur stor del av relationen mellan X och Y som förklaras av M. Sambandet mellan *c - c´* och *ab* har testats statistiskt och under förutsättning att inga data saknas i exempelvis mediatorn, och att beräkningar görs på samma sampel, är i normalfallet *ab* = *c - c´* (MacKinnon, Warsi & Dwyer, 1995).

I jämförelse med att använda testningar av enskilda steg eller differensen mellan *c* och c´ för att undersöka medieringseffekter har produkten *ab* beskrivits som ett mer direkt sätt att testa den indirekta effekten (Fairchild & McQuillin, 2010; MacKinnon m.fl., 2007). Idén om att räkna fram den indirekta effekten genom att multiplicera *a* och *b* presenterades av Sobel redan 1982, och refereras därför ofta till som just Sobel-test (Preacher & Hayes, 2004; Sobel, 1982). Denna metod

har kommit att förespråkas av flera skäl. Bland annat anses *ab* vara mer flexibelt än $c - c'$ då det kan användas i flera olika sammanhang, även när modellen innefattar flera mediatorer (Fairchild & McQuillin, 2010). Jämförelser mellan olika modeller visar dessutom att produkten av *ab* är en betydligt mindre konservativ metod än analyser av enskilda steg, och risken att missa betydande indirekta effekter (typ-II-fel) som faktiskt förekommer är därmed mindre (MacKinnon, Lockwood, Hoffman, West & Sheets, 2002). Mot bakgrund av omfattande tester förespråkas alltså test som bygger på produkten *ab* (d.v.s. Sobel-test och varianter på detta test). Dock behöver vissa justeringar göras för att hantera problemet att produkten av *ab* oftast inte är normalfördelad, vilket diskuteras senare.

Signifikanstestning av indirekta effekter

Efter att ha presenterat olika sätt att räkna fram ett mått på mediering kvarstår fortfarande frågan om vad denna siffra betyder. Med andra ord behöver styrkan i medieringen kunna värderas. Genom signifikanstestning kan en uppskattning göras om medieringseffekten är större än vad som kan antas vara rimligt att andra ovidkommande variabler skulle ha medfört, det vill säga om slumpen kan ha genererat denna effekt (MacKinnon, 2008). Man vill alltså se huruvida den indirekta effekten är signifikant större än 0. Produkten *ab* kan signifikanstestas genom att dividera värdet med ett estimerat standardfel. Det finns formler att tillämpa för att ta fram det estimerade standardfelet (MacKinnon, 2008) och vilken formel som bör användas beror i korthet på hur korrekt den återger det faktiska standardfelet. Genom att dividera medieringseffekten med standardfelet får vi fram ett kritiskt värde som sedan kan jämföras med framräknade statistiska normer. Ett högt kritiskt värde indikerar att medieringseffekten inte är slumpmässig utan en effekt som överensstämmer med verkligheten (MacKinnon, 2008). Tabeller som bygger på normalfördelning indikerar exempelvis att om det absoluta värdet (den indirekta effekten dividerat med standardfelet) överstiger 1.96 är medieringen statistiskt signifikant (större än 0) på $p < 0.05$ nivå.

Användandet av konfidensintervall

För att undersöka huruvida den indirekta effekten är statistiskt signifikant kan man även använda konfidensintervall. Användandet av konfidensintervall förespråkas ofta i olika forskningssammanhang som en modell för att bedöma och diskutera statistisk signifikans på grund av att dessa tar mätfelen i beaktande direkt och därför ger en mer realistisk bild av resultatet än när ett enskilt värde används (MacKinnon, 2008). Genom att använda standardfel kan den övre respektive lägre gränsen för konfidensintervallet skapas utifrån en vald nivå av statistisk signifikans (exempelvis $p<.05$). I normala fall beräknas konfidensintervallet så att intervallen på båda sidor om värdet är lika stora, så kallat symmetriskt konfidensintervall. För att hantera förekomsten av icke normalfördelade medieringseffekter (produkten *ab*) förespråkas användandet av asymmetriska konfidensintervall, som till skillnad från symmetriska intervall är olika stora över respektive under värdet (MacKinnon, 2008). I så kallade simuleringsstudier har man jämfört symmetriska och asymmetriska konfidensintervall och då sett att användandet av asymmetriska konfidensintervall gav ökad power (MacKinnon, Lockwood, Hoffman, West & Sheets, 2002). För en mer ingående beskrivning av hur standardfel beräknas och tillämpas för att räkna fram statistisk signifikans, konfidensintervall och effektstorlekar hänvisas till annan fördjupningslitteratur (MacKinnon, 2008).

Fördelningen av *ab*

Det har visat sig att produkten av två slumpvisa och normalfördelade variabler oftast inte är normalfördelad. Detta har betydande implikationer för medieringsanalyser som bygger på att *a* och *b* multipliceras. Även om fördelningen av *a* och *b* var för sig är normal är alltså inte nödvändigtvis *ab* normalfördelad (MacKinnon m.fl., 2002). I stället har det visat sig att fördelningen av produkten ofta är mycket positivt skev. Samtidigt bygger signifikanstestning i normalfallet på att materialet är normalfördelat. Detta leder till kraftigt försämrad power, eller med andra ord en ökad risk att missa

relevanta medieringseffekter (Preacher & Hayes, 2004). Detsamma gäller när konfidensintervall används för att undersöka om den indirekta effekten är statistiskt signifikant (faller inom konfidensintervallet). Om symmetriska konfidensintervall används och produkten av koefficienterna *ab* inte är normalfördelad leder detta till en betydligt försämrad power och därigenom ökad risk för typ II-fel, det vill säga en ökad risk att felaktigt anta att mediering ej förekommer (MacKinnon m.fl, 2007).

Att slumpa fram nya sampel

En annan metod för att hantera problemet med icke normalfördelade medieringseffekter bygger på att slumpa fram nya sampel, så kallad resampling. Resampling som metod utvecklades för att hantera situationer där den exakta formeln för beräkning av olika statistiska värden saknades eller var oklar, när samplet var för litet för att ge ett tillräckligt statistiskt underlag för beräkningarna eller när fördelningen av data i samplet kunde föranleda problem i analyserna (MacKinnon, 2008). Generellt är fördelen med resampling att man slipper göra lika många antaganden om exempelvis fördelningen av data, då man i stället genererar en utökad databas där man empiriskt kan fastställa detta. Metoder som resamplar den existerande mängden data kräver betydande datorkraft. Av denna anledning har utvecklingen inom området gått framåt kraftigt under de senaste decennierna. Det finns en mängd olika sätt att resampla sitt material som skiljer sig från varandra i hur det nya utökade samplet genereras (MacKinnon, 2008). Att resampla data lämpar sig särskilt väl i just medieringsanalyser, då denna metod resulterar i både tillfredsställande power och mer adekvata nivåer av typ I-fel. Skälet till detta är att resampling inte bygger på antaganden om normalfördelade medieringseffekter, vilket medfört problem i tidigare analysmodeller baserade på enbart det existerande samplet.

Bootstrap-resampling

Så kallad bootstrap-resampling, eller bootstrapping, har rekommenderats för resampling vid medieringsanalyser (MacKinnon m.fl., 2007; Preacher & Hayes, 2004) och är i dag en relativt vanlig metod för att undersöka indirekta effekter. Tillvägagångssättet för att generera ett bootstrap-resampel kan illustreras med ett enkelt exempel. I en studie finns tillgång till data från 50 deltagare, det så kallade originalsamplet. Målsättningen med resampling är att skapa ett visst antal, låt säga 1000, nya sampel med 50 deltagare baserat på originalsamplet. Detta görs genom att slumpvis välja ut först en deltagare, exempelvis nummer 7, och sedan återföra denna till databasen. Därefter väljs ytterligare en deltagare, låt säga nummer 32, vilken sedan också återförs. Detta fortsätter till dess att ett nytt sampel, ett resampel, med 50 deltagare har genererats. Eftersom deltagarna hela tiden återförs efter att ha registrerats i det nya resamplet kan det hända att samma deltagare väljs igen. Slumpen styr hur många gånger en deltagare ingår i det nya resamplet, och på så sätt kommer varje resampel att bli unikt. Baserat på de 1000 resampel som har genererats kan man för varje enskilt resampel räkna fram nödvändiga statistiska parametrar, däribland korrelationerna *a* och *b*. Medelvärdet för en av korrelationerna, baserat på 1000 resampel, benämns den bootstrap-estimerade korrelationskoefficienten. På motsvarande sätt utgör standardavvikelsen det bootstrap-estimerade standardfelet för korrelationskoefficienten. Detta tillvägagångssätt används för att ta fram produkten av koefficienterna *a* och *b* samt standardfelet för medieringseffekten.

De 1000 resampel som genererats används också till att räkna fram konfidensintervall. Det finns flera olika sätt att göra detta på, men här diskuteras den enklaste och mest logiska varianten, bootstrap-percentiler. För att identifiera gränsvärden som motsvarar $p<0.05$, eller med andra ord 95 procents konfidensintervall, sorteras de 1000 estimerade värdena från lägst till högst. Den lägre gränsen i konfidensintervallet utgörs av det 25:e värdet och den högre gränsen av det 976:e värdet (MacKinnon, 2008).

I och med att varje resampel bygger på en slumpmässig sammansättning av observationer från originalsamplet kan resultaten variera något. Med andra ord kan samma analyser ge något varierande konfidensintervall, vilket kan skapa en viss förvirring. Variationen förefaller minska med en ökad mängd resampel och rekommendationerna är därför att slutliga analyser bör bygga på 5000 resampel även om preliminära analyser kan göras med 1000 resampel (Preacher & Hayes, 2008).

Ett flertal olika program har utvecklats eller anpassats för att kunna göra medieringsanalyser baserade på bootstrap-resampel, exempelvis AMOS, Mplus, EQS (MacKinnon, 2008). För SPSS och SAS finns i dag makron som tar fram konfidensintervall för analys av medieringseffekter (Preacher & Hayes, 2004; 2008), som har gjorts tillgängliga på internet.

Även om en stor mängd studier på empiriska grunder pekar i riktning mot att resampling är en bättre modell för medieringsanalys än mer traditionella metoder som enbart baseras på det insamlade materialet bör även möjliga nackdelar och risker framhållas, exempelvis att problem i ett visst dataset riskerar att förstärkas som ett resultat av resampling eller att ett litet sampel kan användas för att generera oproportionerligt mycket analysresultat. Dock förefaller det råda konsensus avseende användbarheten av resampel-metoder. Framför allt för att detta möjliggör analyser av indirekta effekter med tillräcklig power även i studier av begränsad storlek, vilket är av avgörande betydelse för att implementera medieringsanalyser i klinisk behandlingsforskning där storleken på samplet oftast är relativt litet (MacKinnon, 2008).

Power i mediatormodeller

Generellt har medieringsanalyser haft lägre power än analyser av totaleffekter, vilket bland annat har att göra med den indirekta effekten genom värdering av de enskilda koefficienterna i modellen (*a* och *b*). På senare tid har dock metodstudier visat att den traditionellt välanvända metoden att utvärdera enskilda steg i medieringsmodel-

len har alltför låg power för att kunna tillämpas i små och medelstora sampel. Som en illustration av detta har påvisats att en liten medieringseffekt som analyseras utifrån de enskilda stegen kräver 20886 deltagare för att uppnå en power på 0.80 (MacKinnon m.fl., 2007). Metodutvecklingen på senare tid har som beskrivits tidigare mynnat ut i andra metoder som uppvisar bättre balans mellan typ I- och typ II-fel, med en power som bättre lämpar sig för studier baserade på små och medelstora sampel, vilket vanligen förekommer i klinisk behandlingsforskning. Framför allt har metoder som bygger på produkten av koefficienterna *a* och *b* och där man tillämpar asymmetriska konfidensintervall baserade på bootstrap-resampel visat sig ge den bästa kombinationen av en hög power och en rimlig nivå av typ I-fel. Till följd av detta rekommenderas detta tillvägagångssätt i analyserna av indirekta effekter, exempelvis i klinisk behandlingsforskning (Fairchild & McQuillin, 2010; MacKinnon m.fl., 2007; Preacher & Hayes, 2004).

Effektstorlek i mediatormodeller

Precis som i andra sammanhang där signifikanstestning används för att värdera styrkan i olika samband är problemet även vid bedömningen av medieringseffekten att resultatet beror på samplets storlek. Små och relativt obetydliga effekter kan framstå som statistiskt signifikanta om samplet är stort, och betydande medieringseffekter kan komma att negligeras därför att samplets ringa storlek bidragit till att resultaten inte bedöms som statistiskt signifikanta (Cohen, 1988; MacKinnon, 2008). Det är därför viktigt att förhålla sig nyanserat till betydelsen av signifikanstestningar och även diskutera andra sätt att värdera styrkan i medieringseffekter.

Syftet med att ta fram effektstorleksmått som är oberoende av samplets storlek är bland annat att kunna jämföra effekter mellan olika studier och att kunna kombinera dessa i metaanalyser. Avseende medieringsanalyser finns ett antal olika förslag på effektstorleksmått, men någon konsensus eller några tydliga riktlinjer för användandet av dessa har ännu inte formulerats (MacKinnon, 2008). Dessutom

saknas fortfarande alternativ för bedömning av effektstorlekar i mer komplexa modeller med flera mediatorer (Fairchild & McQuillin, 2010). I huvudsak finns två olika typer av effektstorleksmått: dels de som bygger på en värdering av enskilda delar av modellen, dels de som syftar till att bedöma styrkan i hela den indirekta effekten.

Effektstorlek för enskilda delar av medieringsmodellen

Att mäta effektstorleken på enskilda delar av medieringsmodellen kan ha sina tydliga poänger. Detta ger möjligheter att mer noggrant bedöma vilka delar av modellen som är stabila och var i modellen det finns tveksamheter, exempelvis kan styrkan i *a* vara betydligt större än i *b*. Den enklaste och sannolikt vanligaste metoden för att mäta samband mellan två variabler är den bivariata korrelationskoefficienten (*r*). Således kan *r* användas som ett effektstorleksmått för att bedöma styrkan i sambandet mellan X och M (*a*). Fördelen med detta är, förutom enkelheten, att Cohens (1988) kriterier för små (0.1), medelstora (0.3) och stora effekter (0.5) går att tillämpa. I linje med detta kan partiella korrelationer tillämpas för att bedöma styrkan i sambanden mellan X och Y respektive M och Y. Här vill man lyfta ut effekten av antingen X eller M för att på så sätt få en tydligare bild av sambandet med Y, när den ena prediktorn kontrolleras för. På detta sätt kan alltså styrkan i *b* testas genom att i en partiell korrelation mellan M och Y kontrollera för X. Motsvarande partiella korrelation där X får predicera Y med kontroll för M ger information om styrkan i den direkta effekten, eller *c´*.

På motsvarande sätt kan även regressionsanalyser användas för att undersöka styrkan i sambanden, vid kontroll för den variabel man inte vill inkludera. Den standardiserade regressionskoefficienten representerar ett värde som påvisar förändringen i beroendevariabeln när den oberoende variabeln förändras en standardavvikelse. Exempelvis motsvarar en standardiserad *b*-koefficient förändringen i Y vid en standardavvikelses förändring i M, vid kontroll för effekten av X.

Effektstorlekar för hela medieringsmodellen

Till skillnad från mått på enskilda delar av medieringsmodellen finns ett antal olika sätt att värdera styrkan på hela den indirekta effekten, det vill säga produkten av *ab*. Ett vanligt sätt att belysa storleken på den indirekta effekten är att ställa denna i relation till den totala effekten, eller med statistiska termer *ab*/*c*. Med ett sådant förfaringssätt kan resultaten i en medieringsanalys beskrivas i termer av proportion mediering, exempelvis uttryckt som att medieringen svarar för 25 procent av relationen mellan den oberoende och den beroende variabeln. Det kan vara värt att notera den svårighet i tolkningen som kan uppkomma när så kallad inkonsistent mediering förekommer, det vill säga när X och M har olika inverkan på Y, vilket kan medföra att proportionen mediering får ett extremt värde (över 1 eller negativt) (MacKinnon, 2008). En annan variant som förekommer är att den indirekta effekten relateras till den direkta effekten, *ab*/*c´*. Man får då i stället ett mått som avspeglar den indirekta effekten i relation till den direkta effekten.

Antaganden för mediatormodeller

Formulering och analys av medieringsmodeller bygger på flera antaganden som är mer eller mindre möjliga att testa (Fairchild & McQuillin, 2010; MacKinnon m.fl., 2007). En del av dessa är av statistisk karaktär medan andra är av mer konceptuell art. I huvudsak överensstämmer de statistiska antagandena med dem som gäller för vanliga regressionsanalyser (Fairchild & McQuillin, 2010).

Var och en av de regressionsanalyser som medieringsmodellen bygger på utgår från linjära samband, så att en viss förändring i X leder till samma förändring i Y oavsett var på X-axeln detta sker. Dessutom finns ett antagande i enkla mediatormodeller (en mediator) om att relationen mellan X och M är additiv. Med andra ord bör det inte finnas en interaktionseffekt mellan X och M. En interaktionseffekt (XM) betyder i praktiken att relationen mellan M och Y är olika för olika nivåer av X, samt att relationen mellan X och Y varierar beroende på värdet av M. Detta antagande går att testa genom att inkludera

en XM-interaktionsvariabel i modellen (MacKinnon, 2008).

Vidare antas att det inte finns några mätfel till följd av att relevanta variabler utelämnats ur analysen. Exempelvis kan man tänka sig att en annan betydelsefull mediator existerar men inte mäts och därigenom bidrar till att försvåra tolkningen av resultaten. Detta antagande är i praktiken svårt att leva upp till, och framför allt svårt att testa, vilket bör mana till försiktighet i tolkningen av resultaten.

I korthet kan sägas att felvariansen, eller residualerna i de olika ekvationer som medieringsanalyserna bygger på (se ekvationerna 2–4 ovan), inte får korrelera med varandra, inte får korrelera med prediktorvariabeln i respektive ekvation, inte får variera för mycket vid olika nivåer av prediktorvariabeln och dessutom bör vara normalfördelade. När flera ekvationer kombineras bör dessutom felvariansen vara oberoende mellan de olika ekvationerna. Exempelvis finns tillfällen när felvariansen som förekommer i olika ekvationer korrelerar till följd av att de båda är kopplade till en annan, betydande, variabel som utelämnats ur analysen men som påverkar både M och Y (MacKinnon, 2008).

Utöver dessa statistiska förutsättningar bygger medieringsmodellen även på antagandet om att sekvensen av korrelationer är korrekt (X till M till Y och inte Y till M till X), vilket går att testa genom att undersöka den indirekta effekten av Y på X genom M, det vill säga en spegelvänd modell, vilken då inte bör bli signifikant. Dessutom antas att det inte finns ett reciprokt samband mellan M och Y.

Liksom andra typer av analyser bygger även medieringsmodeller på att de instrument som används har acceptabla statistiska egenskaper då mätfel i den föreslagna mediatorn kan leda till felaktiga slutsatser om relationen mellan M och Y. Ofta lyfts vikten av att säkerställa reliabiliteten och validiteten i de mått som används fram, men detta är inte nödvändigtvis tillräckligt. Även om bra utfallsmått existerar innebär forskning kring medieringseffekter att man behöver utveckla adekvata processmått (Kazdin, 2007). Utifrån hypoteser om vilka faktorer som kan tänkas utgöra relevanta mediatorer kan exempelvis faktor- och regressionsanalyser användas för att ta

fram en kombination av item (frågor) med adekvat reliabilitet och validitet. Dock vet man fortfarande inte huruvida instrumentet har förmågan att fånga förekomsten av indirekta effekter, det vill säga kombinationen av relationerna X och M respektive M och Y. Det är således inte tillräckligt att undersöka reliabilitet och validitet för att säkerställa om ett utvecklat processinstrument har tillfredsställande statistiska egenskaper för ändamålet, då själva syftet med instrumentet är att värdera länken mellan X och Y (produkten *ab*). Därför bör även analyser av indirekta effekter ingå i den statistiska utvärderingen av processmått. Analyser av data inhämtade vid ett mättillfälle utgör dock inte grunden för en medieringsanalys, utan syftar till att undersöka om instrumentet på statistiska grunder kan antas fungera som ett adekvat mått på mediering när detta förekommer i longitudinella eller experimentella studier.

Eftersom flera av dessa antaganden är svåra, för att inte säga omöjliga, att testa, och då även några av dem framstår som sannolika (förekomst av vissa mätfel), kan man argumentera för en generell försiktighet i tolkningen av resultat från medieringsanalyser. Detta innebär dock inte ett förespråkande av en alltför konservativ hållning i bedömningen av data. Däremot bör resultaten av medieringsanalysen sättas in i ett sammanhang bestående av flera andra studier för att se till det mönster som resultaten från dessa bildar. Detta resonemang talar också för användandet av flera olika typer av studier. Exempelvis kan resultat från en kvalitativ studie eller data från biologisk grundforskning bidra till att förstärka validiteten i resonemanget kring en möjlig mediator (MacKinnon m.fl, 2007).

Exempel på mediatoranalys i behandlingsforskning med olika design

I jakten på mediatorer är framför allt behandlingsstudier av central betydelse, och dessa kan av varierande skäl ha olika typer av design. Randomiserade kontrollerade studier utgör sannolikt den mest lämpliga formen av design för att undersöka mediering (Kazdin, 2007). Skälet är att vi manipulerar en variabel och därigenom kan

jämföra betydelsen av olika lägen på X, både avseende förändringar i M och effekter på Y. Dessutom är det en logisk tidsföljd i mätningarna där man kan göra flera datainsamlingar under behandlingens gång och därigenom helt eller delvis tillmötesgå tidslinjekriteriet. Man kan möjligen hävda att just randomiseringen, det vill säga den direkta manipulationen av X, utgör en mycket viktig förutsättning för att man i tolkningarna av data ska kunna tala om mediering.

I en randomiserad kontrollerad studie användes produkten av koefficienterna och bootstrap-resampling för att utforska indirekta effekter av psykologisk behandling vid långvarig smärta (Wicksell, Olsson, & Hayes, 2010). Prediktorvariabeln (X) utgjordes av interventionsformen och hade två lägen, Acceptance and Commitment Therapy (ACT) i kombination med sedvanlig behandling eller enbart sedvanlig behandling. Som utfallsmått (Y) användes bland annat smärtrelaterade begränsningar. Flera olika potentiella mediatorer jämfördes för att undersöka om den indirekta effekten var specifik, exempelvis smärtintensitet, rörelserädsla och self-efficacy. Utifrån behandlingens teoretiska ram var hypotesen att de indirekta effekterna skulle vara större i variabeln psykologisk flexibilitet (förmågan att agera i linje med livsvärden och långsiktiga mål i närvaro av smärta) än i övriga potentiella mediatorer. Mätningar i M och Y gjordes vid tre tillfällen, före och efter behandling samt fyra månader efter avslutad behandling. I medieringsanalyserna användes eftermätningen för M och differensen mellan för- och uppföljningsmätningen för Y. För beräkning av konfidensintervall avseende produkten av *ab* användes 3000 resampel. Resultaten från studien visade ett mönster där enbart psykologisk flexibilitet uppvisade signifikanta indirekta effekter, vilket innebär att resultaten uppfyllde kriteriet för specificitet i mediatorn. Däremot fanns oklarheter med avseende på tidslinjekriteriet, vilket behöver undersökas noggrannare i framtida studier. Då eftermätningarna uppvisade förändringar i både M och Y är det omöjligt att säkerställa att effekterna på mediatorn föregått förändringen i utfallsvariabeln.

I en studie för att undersöka förändringsprocesser hos personer

med tinnitus som genomgick en behandling baserad på acceptansstrategier gjordes mätningar av mediatorn och utfallsvariabeln vid varje session (Hesser, Westin, Hayes & Andersson, 2009). Ett av studiens huvudsakliga syften var att säkerställa att effekten i M (bland annat acceptans) föregick förändringen i Y (tinnitussymtom), det vill säga tillmötesgå tidslinjekriteriet. I korthet användes regressionsanalyser för att undersöka om tidiga effekter i M (session 2) predicerade senare effekter i Y (6 månader efter avslutad behandling). Man konstaterade att inga signifikanta förändringar i Y fanns vid session 2, däremot vid session 4. Detta föranledde att mätningarna i M och Y vid session 2 användes som prediktorer i regressionsanalyserna för att undersöka hur dessa förklarade senare effekter på Y. Eftersom man ansåg det tänkbart att tidiga förändringar i Y ändå kunde påverka senare effekter av Y kontrollerade man för de tidiga förändringarna på Y. Hierarkiska regressionsanalyser visade att tidiga förändringar i flera av de hypotetiska mediatorerna förklarade en betydande varians i senare effekter på tinnitussymtom, även när tidiga förändringar i symtom kontrollerades för. Med detta upplägg kunde man alltså på ett tillfredsställande sätt möta tidslinjekriteriet. Denna analysmodell medför dock andra svårigheter i tolkningen av möjliga medieringseffekter: dels finns ingen manipulation av X-variabeln, vilket gör det svårt att utvärdera relationen mellan X och M (det vill säga koefficienten *a*), dels saknas underlag för att säkerställa att den medierande effekten är specifik då andra variabler som enligt teorin inte bör mediera effekten av X på Y ej inkluderades i mätningarna.

Dessa exempel illustrerar några olika tillvägagångssätt för att närma sig frågan om medieringseffekter i behandlingsstudier. I klinisk behandlingsforskning måste analysmodellen ofta anpassas till förutsättningarna för studien, exempelvis finns inte alltid möjlighet att randomisera patienter till olika grupper. Valet av design och analysmetod kan därför medföra olika för- och nackdelar med avseende på de olika kriterier för mediering som har diskuterats. En noggrann planering av datainsamlingen är således av stor betydelse för möj-

ligheten att genomföra meningsfulla medieringsanalyser.

Utvärderingen av behandlingsstudier kan göras utifrån många syften. Först, och kanske främst, vill man vanligen se om interventionen har haft någon effekt på utfallsvariablerna. Medieringsanalysen handlar inte primärt om att svara på frågor om behandlingens effektivitet utan syftar till att klargöra länken eller länkarna mellan behandlingen och effekterna i Y. Med andra ord kan sägas att medieringsanalysen handlar om att undersöka hur effekten på Y genererades, eller vilka M som är viktiga för X inverkan på Y. Av den anledningen kan det vara adekvat att bara inkludera de patienter som uppvisar betydande effekter i M respektive Y för att inte riskera att slutsatser om mediering grumlas av att vissa individer inte hade någon glädje av behandlingen (Kazdin, 2007).

Man bör också ha i åtanke att olika behandlingar kan fungera genom olika mediatorer. Relationen mellan X, M och Y kan således se olika ut om vi utvärderar en symtomreducerande intervention eller om vi använder oss av en behandling med fokus på fysisk träning. Det är exempelvis möjligt att en behandling som syftar till ökad livskvalitet genom minskning av nedstämdhet via användandet av antidepressiv medicinering kommer att uppvisa andra medieringseffekter än den intervention som baseras på frekvent fysisk träning. Detta är av stor betydelse för tolkning och presentation av positiva fynd från en medieringsanalys.

Multipel medieringsmodell

Eftersom man i forskningssammanhang är beroende av förenklingar används modeller som tar mer eller mindre hänsyn till den komplexa verklighet från vilken data hämtas. Mer komplexa modeller återspeglar ibland verkligheten bättre. Dock behöver hänsyn tas till de specifika förutsättningar som gäller för dessa modifierade modeller och komplexiteten av analyserna kan också medföra svårigheter att tolka resultaten (MacKinnon m.fl., 2007; Preacher & Hayes, 2008).

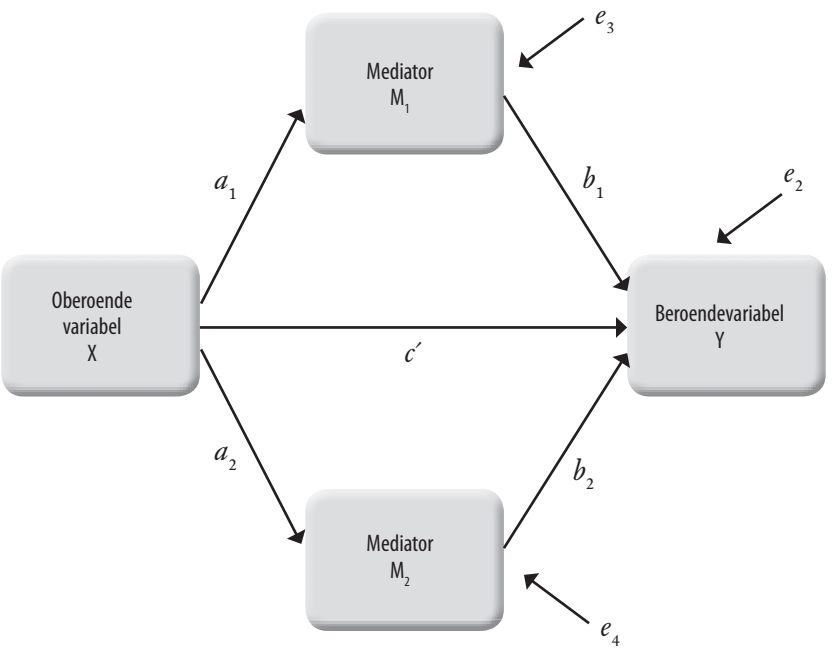
Ett exempel på ovanstående är förekomsten av multipla mediatorer. Det är sannolikt att den medierande effekten av X på Y går via

flera olika mediatorer snarare än via en specifik mediator. På samma sätt kan olika prediktorer påverka samma mediator och en viss mediator kan ha effekter på flera beroendemått. Dessutom är det möjligt att flera mediatorer interagerar för att skapa en indirekt effekt. Baserat på den enkla medieringsmodell som tidigare presenterats kan vi alltså lägga till en eller flera faktorer som bidrar till den indirekta effekten av X på Y.

Figur 15:10 illustrerar en modell med två medierande variabler. Analyser som inkluderar flera potentiella mediatorer i samma modell, det vill säga en multipel medieringsmodell, kan ha flera olika fördelar gentemot en serie av enkla medieringsanalyser som undersöker den indirekta effekten av en mediator i taget. En uppenbar fördel är att man på detta sätt kan bedöma den totala medieringseffekten av flera samverkande mediatorer, den så kallade totala indirekta effekten. Dessutom ges möjlighet att på ett lite annat sätt än i enkla medieringsmodeller värdera betydelsen av varje enskild mediator. Genom att undersöka den specifika indirekta effekten ges ett mått på den specifika variabelns medieringseffekt i närvaro av en ytterligare mediator. En annan aspekt är att en multipel mediatormodell på ett bättre sätt tar hänsyn till antagandet om att inga andra betydande medieringsfaktorer är utelämnade ur modellen, även om denna risk i princip kvarstår så länge den direkta effekten inte är 0 (MacKinnon, 2008). Slutligen medför multipla mediatormodeller ökade möjligheter att jämföra storleken på medieringseffekter genom så kallade kontrastanalyser (Preacher & Hayes, 2008).

I beskrivningen nedan har två medierande faktorer använts för att illustrera skillnader mellan enkla och multipla medieringsmodeller. Detta betecknas, kort och gott, som tvåväga medieringsmodeller. Det går också att tänka sig inblandning av flera mediatorer och man talar då om exempelvis tre- eller fyrväga medieringsmodeller. Modellen som illustreras i figur 15:10 överensstämmer i allt väsentligt med den enkla medieringsmodell som presenterats tidigare. Komplexiteten av modellen blir dock större och det är viktigt att notera skillnaden mellan den enkla och den multipla modellen

med avseende på koefficienten *b*, det vill säga relationen mellan mediatorn och Y. Som modellen visar finns alltså två olika *b*-koefficienter, b_1 och b_2, där b_1 representerar relationen mellan den första mediatorn (M_1) och Y justerat för effekten av b_2 på Y, och b_2 utgör sambandet mellan den andra mediatorn (M_2) och Y med kontroll för den påverkan som b_1 har på Y. Det är således viktigt att notera den avvikande betydelsen av *b* vid multipel mediering jämfört med den enkla medieringsmodellen. En specifik indirekt effekt via M_2 i en multipel medieringsmodell är alltså inte samma sak som den indirekta effekten av variabeln M_2 i en enkel medieringsmodell, efter-



$$Y = i_1 + cX + e_1$$
$$Y = i_2 + c'X + b_1M_1 + b_2M_2 + e_2$$
$$M_1 = i_3 + a_1X + e_3$$
$$M_2 = i_4 + a_2X + e_4$$

Figur 15:10. Path diagram och ekvationer som illustrerar en modell med två medierande variabler, en så kallad multipel medieringsmodell.

som den specifika indirekta effekten representerar den medierande effekten av M_2 i en modell som inkluderar M_1. Anledningen till att medieringseffekten av M_2 blir olika i modeller utan respektive med hänsyn tagen till M_1 är att dessa båda mediatorer sannolikt är korrelerade. I behandlingsforskning blir detta särskilt tydligt. Oftast genomförs en behandling (X) som syftar till att påverka såväl M_1 som M_2, vilket då innebär att dessa variabler har något väsentligt gemensamt, och därigenom högst sannolikt uppvisar ett samband (Preacher & Hayes, 2008).

Analyser av multipla medieringsmodeller bör innefatta två delar. Dels en testning av den totala indirekta effekten, eller med andra ord den kombinerade medieringseffekten av M_1 och M_2, dels en bedömning av den specifika indirekta effekten, eller med andra ord varje mediators enskilda betydelse i den modell som innefattar X, Y och minst en annan mediator. Dock är det inte nödvändigt att den totala indirekta effekten är signifikant för att det ska finnas ett värde i att undersöka de specifika indirekta effekterna, analogt med resonemanget om att det kan förekomma indirekta effekter även när den totala effekten (mellan X och Y) inte uppnår signifikans (Preacher & Hayes, 2008).

I en analys av en multipel medieringsmodell kan man också kontrastera betydelsen av de medierande faktorerna genom att beräkna storleken av de specifika indirekta effekterna, det vill säga det unika bidraget av varje mediator. För en beskrivning av tillvägagångssättet för kontrastanalyser av specifika indirekta medieringseffekter hänvisas till andra artiklar och böcker i ämnet (Preacher & Hayes, 2008; MacKinnon, 2008).

Avancerade mediatormodeller

Som nämnts tidigare kan den grundläggande medieringsmodell som presenterats också tillämpas för att hantera mer komplexa sammanhang. Dock måste vissa anpassningar göras när data innefattar beroendevariabler som är dikotoma, exempelvis arbetslös/i arbete, använder droger/drogfri (MacKinnon, 2007). En noggrann teknisk

beskrivning av hur detta låter sig göras ligger utanför detta kapitels ram, och vi hänvisar till mer avancerad litteratur i ämnet (MacKinnon, 2008; Preacher & Hayes, 2008). I stället följer en kortfattad beskrivning av de olika modellerna.

Analysmodeller baserade på flera nivåer av data, så kallade *linear multilevel modeling*, används allt oftare för att undersöka exempelvis effekter av behandling. I korthet kan detta beskrivas som att man tar hänsyn till de olika nivåer av data där man kan tänka sig att det skapas kluster. Ett tydligt exempel är individer som är indelade i en grupp, vilken utgör en del av en klass, som i sin tur är en av flera på en skola. Om man bortser från möjligheten att data inom respektive kluster (grupp, skola) sannolikt är korrelerade (det finns en samvariation som är relaterad till att data kommer från samma kluster) finns en risk för alltför stora typ I-fel. Detsamma gäller när man vill undersöka medieringseffekter. Om man inte tar hänsyn till att data från grupper kan vara beroende av varandra (samvariera som en effekt av grupptillhörighet), kan det leda till överdrivna slutsatser om mediering. Exempelvis kan man tänka sig att den medierande effekten gäller inom ett visst kluster (en grupp eller en skola) men inte i ett annat (MacKinnon, 2008). Longitudinella studier utgör en avsevärd förbättring avseende möjligheten till medieringsanalyser i jämförelse med tvärsnittsstudier. Viktigt att ha i åtanke är, återigen, tidslinjekriteriet. En databas som innehåller två mättillfällen innebär således fortfarande svårigheter att uttala sig om utvecklingskurvan för M respektive Y. Flera olika analysmodeller har utvecklats för att användas vid longitudinella studier och för fördjupning inom detta område hänvisas till annan litteratur (MacKinnon, 2008).

Integration av moderator- och mediatoranalys

Efter en genomgång av både moderatorer och mediatorer behöver vi nu diskutera möjligheten att dessa effekter förekommer samtidigt, vilket i en komplex verklighet är sannolikt även om man av metodologiska skäl oftast väljer en förenkling som innebär att den

ena eller den andra effekten analyseras som ett isolerat fenomen.

En medieringseffekt som funnits vara av betydelse för relationen mellan en prediktor och en utfallsvariabel kan förstås variera. Exempelvis kan man tänka sig att den medierande effekten av fysisk träning på livskvalitet går via ökat självförtroende, men att denna effekt syns betydligt tydligare hos pojkar än hos flickor. När indirekta effekter förefaller vara påtagligt relaterade till andra variabler finns skäl att misstänka en moderator-mediator-interaktion. Med andra ord verkar den medierande effekten av självförtroende ha modererande faktorer som vi bör ta hänsyn till i vår modell. Det finns en betydande risk att indirekta effekter av relevans för vår förståelse av sambandet mellan X och Y negligeras för att vi missar att ta hänsyn till relevanta moderatorer (Kazdin, 2007). Av denna anledning har många forskare förespråkat att moderatorer och mediatorer kombineras i samma modeller och analyser (MacKinnon, 2008).

Det finns flera olika typer av interaktioner mellan moderatorer och mediatorer men framför allt är moderering av mediation av betydelse för tolkning av data. Detta innebär att den medierande effekten varierar som en effekt av olika nivåer på en moderatorvariabel. En viss behandling kan ha samma effekt på mediatorn, men om relationen mellan mediatorn och beroendevariabeln skiljer sig åt för pojkar och flickor förekommer alltså en skillnad i mediering som kan kopplas till moderatorn kön. Genom att analysera moderatorer av medieringseffekter kan man således identifiera subgrupper av patienter där det förekommer en indirekt effekt, även om den inte är generell för hela samplet. Sådan kunskap kan sedermera ligga till grund för mer sofistikerade hypoteser om mediatorer och moderatorer i exempelvis en viss behandlingsmodell.

Att i statistiska modeller testa mediering med hänsyn också till förekomsten av moderatorer innebär att bättre återspegla den faktiska komplexiteten av sambanden mellan variablerna i modellen. Samtidigt medför denna ökade komplexitet betydande svårigheter med avseende på utvärderingen av effekterna och tolkningen av resultaten. För en mer detaljerad beskrivning av de möjligheter och

svårigheter som moderator-mediator-interaktionen medför hänvisas till MacKinnon (2008).

För att närma sig frågan om medieringseffekter skiljer sig åt för olika subgrupper kan initiala analyser dock göras utan att avancerade statiska modeller av moderator-mediator-interaktion tillämpas. Som i exemplet ovan finns alltså möjligheten att mediering förekommer för pojkar men inte för flickor. Om teorier eller tidigare studier indikerar ett dylikt mönster kan kompletterande analyser av olika potentiella mediatorer där olika subgrupper testas separat ge värdefull information.

Verkningsmekanism

För att ta steget från mediator- till verkningsmekanism rekommenderar Kazdin (2007) ett antal forskningsstrategier.

Använd teori för att vägleda forskningen

Det räcker inte med breda begrepp som psykodynamisk terapi eller kognitiv beteendeterapi utan det behövs specifika begreppsliga modeller som förklarar de processer som ansvarar för den terapeutiska förändringen. Vi behöver mer än test av mediatorer för att förstå mekanismer. Kazdin menar att mediatoranalys kan ha en sorts screeningfunktion för att identifiera potentiella mediatorer att gå vidare med i forskningen.

Inkludera mått på potentiella mediatorer i behandlingsstudier

Mätning av potentiella mediatorer måste planeras in före starten av behandlingsstudien så att man dels har mått som verkligen mäter begreppet av intresse, dels mäter mediatorerna tillräckligt ofta för att ha tidslinjen klar.

Etablera tidslinjen mellan den föreslagna mediatorn och utfallet

Det är av största vikt att kunna visa att den föreslagna mediatorn

förändras före utfallsvariabeln. Kazdin anger att tidslinjen har två krav: 1) mediatorn måste mätas före utfallet och 2) utfallsvariabeln måste också mätas tidigt för att säkerställa att mediatorn verkligen har förändrats före utfallsvariabeln. Idealt mäter man både mediatorn och utfallet vid varje terapisession. Detta gör det möjligt att evaluera mediatorn och utfallsvariabeln och dessutom studera individuella patienters förlopp vad gäller dessa förändringar.

Mät mer än en mediator

Om man bara mäter en mediator, till exempel terapeutisk allians som har gjorts i många studier av psykodynamisk terapi, så kan det bara bli två resultat; antingen finner man ett signifikant samband mellan terapeutisk allians tidigt i behandlingen och utfallet, eller så är sambandet inte signifikant (och då tenderar studien att hamna i byrålådan). Problemet är alltså att om man bara mäter en mediator kan man inte testa om en annan mediator kanske bättre förklarar (medierar) utfallet. Man bör alltså mäta minst två mediatorer; en som stämmer med den teoretiska modell som ligger bakom terapimetoden och en som inte gör det. På så sätt kan man undersöka om den ena mediatorn har ett starkare samband med utfallet än den andra.

Använd designer som kan evaluera mediatorer

Flera olika designer har använts i mediatoranalyser. En vanlig men problematisk design är att man mäter både mediator och utfallsvariabel före och efter behandlingen. Med denna design kan man inte visa tidslinjen och dra slutsatsen att förändring i mediatorn har lett till förändring i utfallsvariabeln. Man kan lika (o)logiskt hävda att förändring i utfallsvariabeln har lett till förändring i mediatorn. Dessutom kan en tredje, okänd variabel ha samband med både mediatorn och utfallsvariabeln. En annan problematisk design är att man mäter utfallsvariabeln före och efter behandlingen men mediatorn under behandlingen. Också denna design missar tidslinjen i och med att utfallsvariabeln inte mäts samtidigt som mediatorn.

Bara för att utfallet inte mäts samtidigt som (eller före) mediatorn betyder det inte att utfallsvariabeln inte har förändrats. En design som kan etablera tidslinjen är att man mäter både mediator och utfallsvariabel före, vid ett tillfälle under och efter behandlingen. Men om man bara mäter en gång under behandlingen är det möjligt att både mediator och utfallsvariabel har förändrats signifikant vid detta tillfälle och då har man fortfarande inte en etablerad tidslinje. Den bästa, men också mest arbetskrävande designen är att man mäter både mediator och utfallsvariabel före, vid varje session under behandlingen och efter avslutad behandling. Med denna design är det möjligt att studera när en signifikant förändring i mediatorn sker i förhållande till förändring i utfallsvariabeln och studera förloppen för varje enskild patient.

Undersök konsistenser över olika typer av studier

Kazdin rekommenderar att olika typer av forskningsstudier görs och att man sammanställer dessa för att undersöka om det finns konsistenta fynd över olika studier. Han nämner till exempel studier från djurlaboratorier, naturalistiska studier, kvalitativa studier och laboratoriestudier av terapeutiska processer. Konsistenta fynd över olika studier gör det mer plausibelt att man har identifierat en hållbar mediator.

Gör en interventionsstudie för att förändra den föreslagna mediatorn/mekanismen

Ovanstående strategier sysslar i grunden med korrelationsdata och för att kunna dra kausalslutsatser behöver man göra experimentella studier. En sådan strategi är att göra en behandlingsstudie där den mediator som kommit fram i de tidigare strategierna varieras över grupper. Man rekryterar alltså patienter med samma diagnos och randomiserar sedan dessa till två betingelser som alla får samma behandling utom med avseende på den aktuella mediatorn. I den ena gruppen (hög) maximerar man denna faktor och i den andra gruppen (låg) minimerar man faktorn. Sedan måste man naturligtvis

mäta mediatorn för att visa att grupperna skiljer sig åt i detta avseende; om inte har man misslyckats med manipulationen av den oberoende variabeln (Kazdin, 2003). Om man sedan får signifikant bättre resultat på utfallsvariabeln i hög- än i låggruppen så har man sannolikt identifierat en verkningsmekanism för den aktuella behandlingsmetoden.

Denna typ av experimentell studie finns, såvitt vi vet, dock ännu inte gjord för någon enda terapimetod, oberoende av teoretisk grund. Det innebär att forskningen har mycket arbete kvar för att man ska kunna tala om verkningsmekanismer på empirisk grund och inte bara som teoretisk modell.

Sammanfattning

Den som vill undersöka moderatorer måste specificera dessa innan datainsamlingen startar eftersom det rör sig om karakteristika som individerna har med sig in i en studie. Dessutom måste man göra en poweranalys (se kapitel 4) så att man har tillräcklig statistisk power när det ursprungliga samplet delas på hälften eller ännu mindre.

Är man intresserad av mediatorer så är det viktigt att mäta minst två mediatorer: en som följer av den teori som ligger bakom behandlingsmetoden och en som är förankrad i en helt annan teori. Dessutom måste man se till att klara tidslinjen, det vill säga att mäta både mediatorer och utfallsmått så ofta att man kan visa förändring i mediatorn innan det sker en förändring i utfallsvariabeln.

Om man har kommit så långt att en mediator för en behandlingsmetod är empiriskt fastställd och man vill undersöka en verkningsmekanism rekommenderas att man manipulerar den funna mediatorn i en experimentell studie där två grupper av patienter får samma behandling, X, men där de får olika omfattning av mediatorn, till exempel M_1 respektive M_0. Om man då finner att M_1-gruppen får signifikant bättre effekt än M_0-gruppen är det ett starkt stöd för att man har funnit en verkningsmekanism för behandlingen ifråga.

Fördjupningslitteratur

Kazdin, A. E. (2007). Mediators and mechanisms of change in psychotherapy research. *Annual Review of Clinical Psychology, 3*, 1-27.

Kraemer, H. C., Wilson, G. T., Fairburn, C. G. & Agras, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry, 59,* 877-883.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York: Psychology Press.

Referenser

Aiken, L. S. & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage Publications.

Baron, R. M. & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1182.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ: Erlbaum.

Fairchild, A. J. & McQuillin, S. D. (2010). Evaluating mediation and moderation effects in school psychology: A presentation of methods and review of current practice. *Journal of School Psychology, 48*, 53–84.

Hesser, H., Westin, V., Hayes, S. C. & Andersson, G. (2009). Client's in-session acceptance and cognitive defusion behaviors in acceptance-based treatment of tinnitus distress. *Behaviour Research and Therapy, 47,* 523–528.

Holmbeck, G.N. (1997). Toward terminological, conceptual, and statistical clarity in the study of mediators and moderators: Examples from the child-clinical and pediatric psychology literatures. *Journal of Consulting and Clinical Psychology, 65*, 599–610.

Kadden, R. M., Cooney, N. L., Getter, H. & Litt, M. D. (1989). Matching alcoholics to coping skills or interactional therapies: Posttreatment results. *Journal of Consulting and Clinical Psychology, 57*, 698–704.

Kazdin, A. E. (2003). *Research design in clinical psychology, 4th ed.* Boston: Allyn & Bacon.

Kazdin, A. E. (2007). Mediators and mechanisms of change in psychotherapy research. *Annual Review of Clinical Psychology, 3*, 1–27.

Kraemer, H. C., Wilson, G. T., Fairburn, C. G. & Agras, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry, 59,* 877–883.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York: Psychology Press.

MacKinnon, D. P., Fairchild, A. J. & Fritz, M. S. (2007). Mediation analysis. *An-*

nual Review of Psychology, 58, 593–614.
MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G. & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7*, 83–104.
MacKinnon, D. P., Warsi, G. & Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research, 30*, 41–62.
Preacher, K. J. & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers, 36*, 717–731.
Preacher, K. J. & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*, 879–891.
Project MATCH Research Group (1998). Matching patients with alcohol disorders to treatments: Clinical implications from Project MATCH. *Journal of Mental Health, 7,* 589–602.
Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In L. S (Ed.), *Sociological Methodology* (pp. 290–312). San Fransisco: Jossey-Bass.
Wicksell R. K., Olsson G. L. & Hayes S. C. (2010). Psychological flexibility as a mediator of improvement in Acceptance and Commitment Therapy for patients with chronic pain following whiplash. *European Journal of Pain, 14*, 1059. e1–1059.e11.
Öst, L-G., Jerremalm, A. & Jansson, L. (1984). Individual response patterns and the effects of different behavioral methods in the treatment of agoraphobia. *Behaviour Research and Therapy*, *22*, 697–707.
Öst, L-G., Jerremalm, A. & Johansson, J. (1981). Individual response patterns and the effects of different behavioral methods in the treatment of social phobia. *Behaviour Research and Therapy, 19*, 1–16.
Öst, L-G., Johansson, J. & Jerremalm, A. (1982). Individual response patterns and the effects of different behavioral methods in the treatment of claustrophobia. *Behaviour Research and Therapy, 20*, 445–460.

Kapitel 16 | Sten Anttila

Tolkning av resultat

En effekt brukar beskrivas som skillnaden mellan vad som hänt efter en intervention och vad som skulle ha hänt om inte interventionen gjorts.

> In an experiment we observe what did happen when people received a treatment. The counterfactual is the knowledge of what would have happened to those same people if they simultaneously had not received treatment. An effect is the difference between what did happen and what would have happened. (Shadish, Cook & Campbell, 2002, s. 5)

Begreppet ”effektstorlek” brukar i sin tur definieras statistiskt som den utsträckning i vilken ett fenomen existerar i en given population.

> Without intending any necessary implication of causality, it is convenient to use the phrase ”effect size” to mean ”the degree to which the phenomenon is present in the population,” or ”the degree to which the null hypothesis is false.” Whatever the manner of representation of a phenomenon in a particular research in the present treatment, the null hypothesis always means that the effect is zero. By the above route, it can now readily be made clear that when the null hypothe-

> sis is false, it is false to some specific degree, i.e. the effect size (ES) is some specific value in the population. The larger this value, the greater the degree to which the phenomenon under study is manifested. (Cohen, 1988, s 9-10)

Nollhypotesen innebär att skillnaden mellan vad som hände och vad som skulle ha hänt är noll om den kausalitet som Shadish med kollegor (2002) skriver om beaktas. När man drar slutsatsen att en given observerad effektstorlek är ”statistiskt signifikant” (t.ex. $p < .05$), så betyder detta i grova drag att den observerade effekten är så pass stor att sannolikheten att den skulle ha observerats, under förutsättningen att nollhypotesen är sann, är mindre än 5 procent.

> The P value is the probability of having observed our data (or more extreme data) when the null hypothesis is true. (Altman 1991, s. 167)

Om effekten däremot inte är statistiskt signifikant så innebär detta att effektstorleken är så pass liten att sannolikheten att den skulle observeras i utvärderingen är fem procent eller mer ($p \geq .05$), under förutsättning att nollhypotesen är sann. Det är dock inte självklart att statistisk signifikans betyder att effekten är stor. Även en liten effekt kan vara statistiskt signifikant om antalet individer som ingår i studien är tillräckligt många.[1]

Syftet med detta kapitel är att visa hur effekt, effektstorlek samt statistisk signifikans kan tolkas inom ramen för en evidensbaserad praktik, där valet står mellan alternativa interventioner och där den kliniska expertisen har att väga samman patientens/klientens preferenser och beteenden med beaktande av klientens/patientens kliniska tillstånd och andra lokala omständigheter samt med hänsyn tagen till evidens från forskningen (Haynes m.fl., 2002). Den randomiserade studien (RCT) är normalt sett den mest tillförlitliga

1 Detta beror på att samplingsfördelningens spridning, standardfelet, minskar då antalet individer ökar (allt annat lika).

studiedesignen vid effektutvärderingar. Det finns emellertid skäl att diskutera några problem förenade med RCT och tolkning av effekter, vilket är kapitlets sista syfte.

Effekt och komparativ effekt

När den skattade effekten av en intervention tolkas inom ramen för en evidensbaserad praktik bör flera frågor beaktas. Det är inte säkert att det kontrafaktiska betyder samma sak i alla olika kliniska sammanhang. Det är ofta rimligt att anta att om interventionen inte genomförts, så skulle en eller flera alternativa interventioner ha genomförts. I vissa fall kan det dessutom vara så att ingen intervention skulle ha genomförts (vilket t.ex. kan vara fallet inom prevention). I ett land kan alternativet till interventionen vara en annan specifik intervention, medan det i ett annat finns tre alternativ och i ett tredje land inget alternativ alls. Kontrollvillkoren kan även variera över tid, eftersom vissa interventioner fasas ut och nya kan tillkomma.

Det som kontrollgruppens deltagare exponeras för är alltså av central betydelse för effekten. Effektstorleken beror ju på hur stor skillnaden är mellan interventionsgruppens och kontrollgruppens resultat. Det är ju inte bara så att ju bättre intervention, desto större blir effekten. Det omvända gäller även. Ju sämre situationen varit för kontrollgruppen, desto större blir effekten (med en och samma intervention). Detta glöms ibland bort. Ibland tar sig detta uttryck i att man endast talar om att en intervention har effekt utan att närmare gå in på vad detta betyder. Det kan även visa sig i att man inte redovisar det kontrollgruppen exponerats för på samma detaljerade sätt som interventionen. Att till exempel endast skriva att kontrollgruppen fick standardbehandling (eng. *treatment as usual*) utan någon närmare specificering innebär att resultaten kan vara svåra att tolka och ibland helt meningslösa.

Antingen får kontrollgruppens deltagare någon typ av alternativ intervention eller så får de ingen egentlig intervention. Om de inte

Tabell 16:1. Typologi över alternativa jämförelser och effekter.

	Kontrollgrupp
Effekt	Ingen intervention: isolerad kontrollgrupp (ibland väntelista)
	Placebointervention
Komparativ effekt	Standardinterventioner: flera alternativ
	Specifik intervention
	Flera specifika interventioner

får någon egentlig intervention så kan man tala om en ”ren” effekt.[2] Om alternativet är placebo, kan man uppskatta hur stor placeboeffekten kan tänkas vara (jfr Hróbjartsson & Gøtzsche, 2004), för att därefter justera resultaten[3].

I övriga fall, när alternativet är en eller flera specifika interventioner eller standardbehandling, handlar effekten alltid om en förbättring i relation till alternativet som är mer eller mindre väldokumenterat, alltså en komparativ effekt (jfr Luce, Kramer, Goodman, Connor, Tunis, Whicher m.fl., 2009). I tabell 16:1 föreslås en typologi över olika jämförelser som kan göras vid effektutvärderingar. Syftet med förslaget är att underlätta tolkandet av effektstorlekar och det motiveras av att det ännu inte finns någon internationellt etablerad och ämnesöverskridande typologi.

I akademisk forskning eller under ett tidigt forskningsskede, när en intervention håller på att utvecklas, så kan målet vara att skatta den effekten av en intervention (jfr Shadish m.fl., 2002, s. 259). Detta innebär att en isolerad kontrollgrupp med ingen intervention eller en placebointervention utgör kontrollalternativen. Inom

2 Om det rör sig om ”ingen” intervention, så krävs det egentligen att man ”isolerar” kontrollgruppen från alla övriga tänkbara influenser (jfr Shadish m.fl., 2002, s. 246–7) – till exempel förhindra deltagarna från att på egen hand söka vård för sina problem. I annat fall är det problematiskt att tala om en ren ”effekt”.

3 Den slutsats Hrobjartsson och Gøtzsche drar är emellertid följande: ”There was no evidence that placebo interventions in general have clinically important effects. A possible small effect on continuous patient-reported outcomes, especially pain, could not be clearly distinguished from bias.”

en evidensbaserad praktik är denna effekt inte alltid lika intressant. Här behövs diskriminerande information till stöd för valet mellan olika alternativa interventioner. Denna jämförande forskning (eng. *comparative effectiveness research*) beskrivs av Manchikanti, Falco, Boswell och Hirsch (2010, s. E24) på följande sätt:

> …the generation and synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat, and monitor a clinical condition or to improve the delivery of care. /…/ to assist consumers, clinicians, purchasers, and policy-makers to make informed decisions that will improve health care at both the individual and population levels.

Den ”rena” effekten är mindre intressant om en komparativ effekt, baserad på direkta jämförelser av de alternativ man har att välja mellan i praktiken, finns tillgänglig. Om denna direkta jämförelse är möjlig kan man tala om direkt evidens (se GRADE Working Group, 2004; Brozek, Akl, Alonso-Coello, Lang, Jaeschke, Williams m.fl., 2009). Om de alternativa interventionerna inte jämförts med varandra utan med andra alternativ, så kan man tala om indirekt evidens (eng. *indirectness of evidens*) vilket innebär att resultaten är mindre tillförlitliga än vid direkt evidens (jfr Glenny, Altman, Song, Sakarovitch, Deeks, D’amico m.fl., 2005). Även standardbehandling som kontrollalternativ kan medföra en indirekt evidens, till exempel om standardbehandling innebär att flera olika interventioner kollapsat och/eller om någon eller all standardbehandling i princip inte finns tillgänglig lokalt där valet görs.

Om standardalternativet till den intervention som utvärderas är ingen behandling, vilket ibland kan vara fallet inom till exempel prevention, så innebär naturligtvis placebokontroll inget problem rörande problemet med indirekt evidens. En liknande situation finns där den utvärderade interventionen är ett nytt komplement till befintliga behandlingar. Här får deltagarna i interventionsgruppen den nya interventionen förutom standardbehandling medan deltagarna

i kontrollgruppen får samma standardbehandling men ett komplement i form av placebo. Inte heller här utgör placebokontroll något problem avseende indirekt evidens.

Sammanfattningsvis, när man tolkar resultaten från en effektutvärdering bör man skilja mellan effekt och komparativ effekt samt bedöma i vilken utsträckning evidensen är direkt eller indirekt.

Statistiska definitioner av effektmått

Skillnaden mellan det faktiska utfallet av interventionen och det hypotetiska utfallet, om interventionen inte gjorts, alltså effekten, kan specificeras statistiskt på flera olika sätt. Detta kan innebära problem då man ska välja mellan alternativen, när man ska tolka resultaten i en given studie samt då man ska väga samman resultaten från flera studier. Det finns därför skäl att använda de standardalternativ som brukar rekommenderas inom det internationella nätverket Cochrane Collaboration. Några av dessa vanligt förekommande statistiska effektmått kommer att beskrivas nedan.

Utgångspunkten brukar vara att dela in utfallen i kontinuerliga respektive binära utfall (jfr Higgins & Green, 2008, s. 250–257; Borenstein, Hedges, Higgins & Rothstein, 2009, s. 17–57). För kontinuerliga utfall brukar medelskillnad[4] (*D*) eller standardiserad medelskillnad (t.ex. Cohens *d*) vara vanliga alternativ. Om utfallen är binära används ofta oddskvot (*OR*), relativ risk (*RR*) eller riskskillnad (*RD*).[5]

4 Termen ”medelvärdesskillnad” är egentligen en bättre översättning, men vi har valt det kortare alternativet ”medelskillnad” eftersom det är kortare och ligger nära den engelska termen (”mean difference”).

5 Ett binärt statistiskt effektmått som blivit allt vanligare är hasardkvoten (hazard ratio, HR): Detta mått används när tiden fram till en händelse är viktig (se Altman, 1991, 365–95).

Tabell 16:2. Kontinuerliga utfall.

	Gruppstorlek	Medelvärden	Standardavvikelse
Intervention	n_I	m_I	sd_I
Kontroll	n_C	m_C	sd_C

Kontinuerliga utfallsmått

Medelskillnaden *D* är förhållandevis lätt att förstå. Man minskar helt enkelt medelvärdet för utfallet i interventionsgruppen med motsvarande medelvärde i kontrollgruppen (ekvation 1 samt tabell 16:2) vid en tidpunkt efter att interventionen genomförts[6].

$$D = m_I - m_C \qquad (1)$$

Om det är eftersträvansvärt (t.ex. god hälsa) att ha ett högt värde, så kommer *D* att vara positivt om interventionen har en bättre effekt än jämförelsealternativet (det som kontrollgruppen exponerats för). Om utfallet å andra sidan avser något man vill undvika (t.ex. sjukdom), så kommer *D* att vara negativt om interventionen har en bättre effekt.

Ett problem med *D* som effektmått är att resultaten från snarlika utfall kan vara svåra att jämföra eftersom resultatet är beroende av vilket instrument och vilken skala som använts vid mätningarna. Ett resultat uppmätt på en skala (t.ex. 1–5) är inte jämförbart med ett resultat på en annan skala (1–20).

Cohens *d* utvecklades för att underlätta analyser avseende statistisk teststyrka (power). För att detta skulle vara praktiskt möjligt krävdes ett mått på effektstorlek som är oberoende av vilka skalor och enheter som ett givet fenomen uppmätts med. Genom att dividera *D* med den poolade standardavvikelsen *s* (ekvation 3 samt

6 Ibland väljer man att jämföra skillnader i förändring från en tidpunkt före interventionen ägt rum till en tidpunkt efter (change scores), mellan interventions- respektive kontrollgrupp.

tabell 16:2) från interventions- respektive kontrollgruppen, så får man effektmått med sådana egenskaper (ekvation 2 samt tabell 16:2).[7]

$$d = \frac{D}{s} \tag{2}$$

…där

$$s = \sqrt{\frac{(n_I - 1)sd_I^{\,2} + (n_C - 1)sd_C^{\,2}}{N - 2}} \tag{3}$$

…och

$$N = n_I + n_C \tag{4}$$

Cohens *d* är alltså inte en medelskillnad rörande en konkret variabel, till exempel antal sjukdagar, utan en skillnad mellan medelvärden där enheten utgörs av standardavvikelser. Att *d* är skaloberoende innebär att ett resultat uttryckt som *d* i en undersökning kan jämföras med ett resultat *d* från en annan undersökning, även om inte identiska mätinstrument och skalor använts. Detta är av stor vikt, om man inom ramen för en evidensbaserad praktik har att välja mellan två alternativa interventioner och direkta jämförelser saknas. Fördelen med *d* är att jämförelser mellan resultat från olika studier underlättas.

I tabell 16:3 illustreras hur medelskillanden *D* och den standardiserade medelskillnaden *d* kan räknas fram. Uppgifterna kommer från en RCT rörande alternativa boendeprogram för hemlösa (vilka båda inkluderar varianter av intensiv case management). Integrated Housing Services (*IHS*) innebär i korthet att boendet är en integrerad del av behandlingen och att vårdgivaren står för bostaden där

7 Hedges *g* är en vanligt förekommande modifiering av *d* för att komma tillrätta med ”small sample bias” där $g = d\ (1 - 3/(4df - 1))$.

Tabell 16:3. Kontinuerliga utfall – andel dagar i stabilt boende per person, uppföljning efter 18 månader (från McHugo, Bebout, Harris, Cleghorn, Herring, Xie m.fl., 2004).

	Gruppstorlek	Medelvärden	Standardavvikelse
Integrated Housing Services (IHS)	54	0,85	0,27
Parallel Housing Services (PHS)	48	0,68	0,40

även personal finns tillgänglig (åtminstone i anslutning till bostaden). Parallel Housing Services (*PHS*) innebär att boende och behandling hålls separata. Deltagande i behandling är alltså inte kopplat till boendet. Bostäderna kommer från den öppna marknaden och det finns ingen personal i anslutning till bostäderna. Resultatet innebär att deltagarna i *IHS* uppvisade 17 procent fler dagar i stabilt boende än de som fått *PHS* (D=0,17). Det kan vara svårt att jämföra detta utfall med ett resultat där stabilt boende mätts på ett något annorlunda sätt. För att underlätta sådana jämförelser kan man använda den standardiserade medelskillnaden (d), vilken här är 0,50.

$$D = 0{,}85 - 0{,}68 = 0{,}17$$

$$s = \sqrt{\frac{(54-1)0{,}27^2 + (48-1)0{,}40^2}{102-2}} = 0{,}337$$

$$d = {}^{0{,}17}\!/_{0{,}337} = 0{,}504 \approx 0{,}50$$

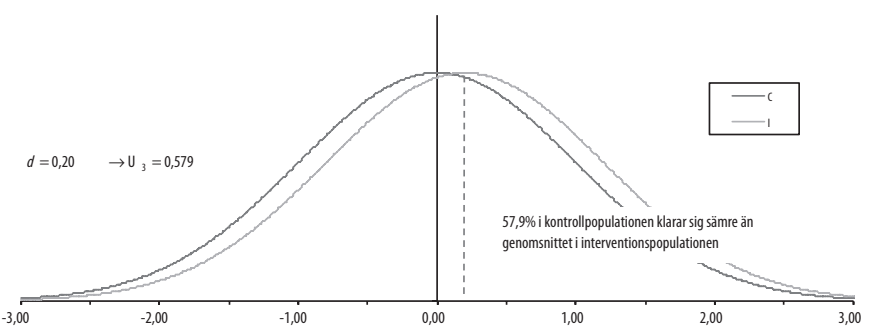
Problemet med d är att den konkreta innebörden för beslutsfattarna inom den evidensbaserade praktiken går förlorad. Vad menas till exempel med att d=0,2 eller d=2,0? Betrakta d som en populationsparameter där de två populationerna (interventionsgruppen respektive kontrollgruppen) är normalfördelade, är lika stora och har samma varians. Under dessa omständigheter kan d tolkas som en skillnad mellan två Z-värden i den standardiserade normalfördelningen. Det-

ta betyder att man med hjälp av den kumulativa normalfördelningen kan ge *d* en mer praktiknära innebörd. Cohen utvecklade i detta sammanhang ett index kallat *U* och det kanske mest användbara U_3.

> If we maintain the assumption that the populations being compared are normal and with equal variability, and conceive them further as equally numerous, it is possible to define a measure of nonoverlap (*U*) associated with *d* which are intuitively compelling and meaningful… as third measure of overlap, U_3, we take the percentage of the A population which the upper half of cases of the B population exceeds. When *d*=0, U_3=50.0%... [when] *d*=2.0… the upper half of the B population exceeds A population, so that U_3=97.7%. (Cohen, 1988, s. 21–3)

I Cohens tabell 2.2.1 (1988, s. 22) framgår att varje värde för U_3 motsvarar en andel som *Z*-värdet ger i den kumulativa normalfördelningen. Med hjälp av U_3 så kan man i viss mån konkretisera vad en given effektstorlek kan betyda. I figur 16:1 nedan illustreras vad en effektstorlekt *d*=0,20 kan betyda.

Under givna förutsättningar betyder en effektstorlek *d* på 0,20 att 57,9 procent i kontrollpopulation klarar sig sämre än den bättre hälften inom interventionspopulationen. Arean under den röda



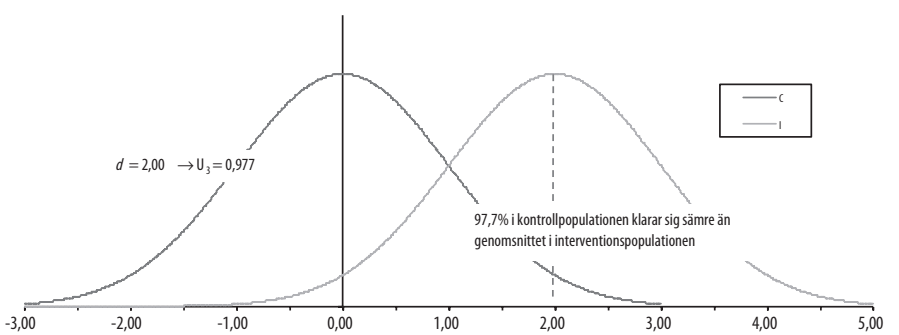
Figur 16:1. Illustration på Cohens *d* på 0,20.

kurvan (kontrollpopulationen) till vänster om den streckade linjen (medianen för interventionspopulationen) är 57,9 procent av den totala arean under den röda kurvan. Att tillhöra interventionspopulationen innebär alltså en förbättring på 8 procentenheter jämfört med kontrollpopulationen (avseende rangordning).

I figur 16:2 ges motsvarande illustration av en effektstorlek *d* på 2,00. Hela 97,7 procent i kontrollpopulationen klarar sig sämre än den bättre hälften i interventionspopulationen. I detta fall är motsvarande förbättring 48 procentenheter.

Om man återknyter till tolkningar inom ramen för en evidensbaserad praktik är det förstås av avgörande betydelse om *d* avser en effekt eller en komparativ effekt. Om det handlar om en komparativ effekt är det av central betydelse att veta exakt vad kontrollgruppens deltagare exponeras för. Om det rör sig om en komparativ effekt i figur 16:1 och en effekt i figur 16:2, så kan resultatet i figur 16:1 mycket väl vara ett starkare argument för att använda interventionen än resultatet i figur 16:2. Detta skulle vara fallet om resultatet i figur 16:1 handlade om en direkt jämförelse och om kontrollalternativet i denna figur medförde en lika stor effekt som interventionen i figur 16:2 i tredje utvärdering.

Skillnaden mellan ”ren” effekt och komparativ effekt påverkar beräkningar av statistisk teststyrka (power). Man bör till exempel



Figur 16:2. Illustration på Cohens *d* på 2,00.

Tabell 16:4. Binära utfall.

	Händelser	Ej händelser	Summa
Intervention	A	B	n_1
Kontroll	C	D	n_2

utgå från att en förväntad komparativ effekt är mindre än en motsvarande ”ren” effekt. Man bör knappast utgå från att effekten, enligt Cohens (1988, s. 24–7) indelning, är stor eller medium när antalet deltagare bestäms (vilket inte är ovanligt). I stället bör man utgå från att effekten är liten och basera beräkningarna på tidigare utvärderingar rörande det aktuella alternativet.[8]

Binära utfallsmått

Ganska ofta förekommer binära utfallsmått då interventioner utvärderas. Antingen är man frisk efter en behandling eller så är man fortfarande sjuk. Antingen har händelsen inträffat A respektive C antal gånger (antalet tillfrisknade personer) eller så har den inte inträffat B respektive D antal gånger (antal sjuka personer).

Det effektmått som är enklast att tolka är riskskillnad (*RD*). Detta effektmått räknas fram genom att man subtraherar andelen där händelsen inträffat (t.ex. att ha ett stabilt boende som före detta hemlös) inom kontrollgruppen med motsvarande andel inom interventionsgruppen (ekvation 5 samt tabell 16:4).

$$RD = \frac{A}{n_1} - \frac{C}{n_2} \qquad (5)$$

En *RD* på 0,10 innebär att andelen i stabilt boende i interventionsgruppen är 10 procent fler än dem i kontrollgruppen. Ett sätt att

8 Ett specialfall av frågeställning av intresse här rör om interventionen, med avseende på ett givet utfall (t.ex. biverkningar) i alla fall inte är sämre än kontrollalternativet, en så kallad ”non-inferiority” studie (Piaggio, Elbourne, Altman, Pocock & Evans for the CONSORT group, 2006). I detta fall krävs normalt ett betydligt större antal individer än vid standardfrågeställningar samt att man definierat en klinisk gräns för vad som uppfattas som acceptabel skada.

ytterligare konkretisera detta är att räkna fram hur många som behöver få interventionen för att åtminstone en person ska uppnå det önskade tillståndet, till exempel att som före detta hemlös klara av ett stabilt boende. Detta får man fram genom att dividera ett med *RD* – detta brukar kallas ”number needed to treat” eller NNT (ekvation 6). Om *RD* är 10 räcker med att 10 får interventionen för att åtminstone en person ska tillfriskna.

$$NNT = \frac{1}{RD} \qquad (6)^{9}$$

Det kanske vanligaste binära effektmåttet är emellertid oddskvoten (eng. *odds ratio, OR*). *OR* får man fram i två steg. Det första steget innebär att man räknar fram oddsen för händelsen i interventionsgruppen respektive kontrollgruppen genom att dividera antalet händelser med antalet icke-händelser (t.ex. antalet friska dividerat med antalet sjuka). Därefter dividerar man oddsen för händelsen interventionsgruppen med oddsen för kontrollgruppen (ekvation 5 samt tabell 16:4)[10].

$$OR = \frac{A/B}{C/D} \qquad (7)$$

En oddskvot på 1,00 innebär att det inte finns någon skillnad mellan oddsen i de två grupperna. En oddskvot på 0,5 innebär att oddsen är hälften så stora för personer inom interventionsgruppen att erfara händelsen som oddsen i kontrollgruppen.

Relativ risk (*RR* för risk-ratio) är ytterligare ett vanligt förekommande effektmått. I detta fall dividerar man risken i interventionsgruppen med risken i kontrollgruppen. En *RR* på 1,00 innebär att risken är lika i de båda grupperna, medan en *RR* på 0,05 innebär att risken är en tjugondel av risken inom kontrollgruppen.

9 Det finns även motsvarande ”number needed to harm” (NNH).
10 Ibland används följande förenklade formel: $OR = \frac{A \cdot D}{B \cdot C}$

Tabell 16:5. Binära utfall – stabilt boende (modifierat från Morse, Calsyn, an Klinkenberg, Helminiak, Wolff, Drake m.fl., 2006).

	Händelser	Ej händelser	Summa
Assertive Community Treatment (ACT)	25	29	54
Ordinarie mental- och missbruksvård	19	30	49

$$RR = \frac{A/n_1}{C/n_2} \tag{8}$$

I tabell 16:5 och i tabellens anslutning illustreras hur man kan beräkna olika statistiska effektmått då utfallet är binärt. Uppgifterna kommer från en RCT där Assertive Community Treatment (ACT) jämförts med ordinarie mental- och missbruksvård för hemlösa. ACT innebär i korthet att multidisciplinära team arbetar intensivt med klienterna i samhället där klienterna bor (och inte på institution). Tillgängligheten till teamen är maximal, alltså 24 timmar per dygn under veckans alla dagar, och antalet klienter per medlem i teamet är lågt (ca 10).

$$RR = \frac{25/54}{19/49} = \frac{0{,}46}{0{,}39} = 1{,}18$$

$$OR = \frac{25/29}{19/30} = \frac{0{,}86}{0{,}63} = 1{,}37$$

$$RD = \frac{19}{49} - \frac{25}{54} = 0{,}39 - 0{,}46 = -0{,}07 \rightarrow RD = 7\%$$

$$NNT = \frac{1}{0{,}07} = 14{,}28 \rightarrow NNT \approx 14$$

I detta fall är *RR* 1,18, alltså 18 procent högre risk att de som fått *ACT* uppvisar ett stabilt boende vid 12-månadersuppföljningen än att de som fått ordinarie vård gör det. *OR* är något högre, 1,37. *RD*

innebär att sju procent fler i *ACT*-gruppen har ett stabilt boende jämfört med kontrollgruppens deltagare, vilket innebär att man måste ge *ACT* till minst 14 personer för att man sannolikt ska få åtminstone en person mer i stabilt boende jämfört med ordinarie vård. *RD* och *NNT* är lätta att förstå och är användbara i samhällsekonomiska sammanhang, men ger låga värden då andelen händelser är liten (ovanliga sjukdomar eller ovanliga förbättringar). *RR* kan vara något enklare att förstå än *OR*, men är något mindre känslig för ovanliga händelser. Om skillnaden i utfall mellan interventionsgruppen och kontrollgruppen dessutom är stor så ger OR större utslag än RR.

Även för binära utfallsmått är det av central betydelse om det rör sig om en effekt eller en komparativ effekt. Om interventionsgruppen fått en given behandling och kontrollgruppen med säkerhet inte fått någon behandling alls, så betyder en OR på 0,5 något annat än om kontrollgruppen fått en alternativ behandling. Samma sak gäller förstås även RD, NNT och RR. Speciellt förrädiskt kan det vara med NNT, antalet personer som måste behandlas för att åtminstone en ska tillfriskna. Det handlar om NNT i relation till en alternativ behandling och inte om en specifik behandling tagen för sig. Även i detta fall bör man beakta skillnaden mellan komparativ effekt och ”ren” effekt då den statistiska teststyrkan beräknas.

Tolkning av statistisk signifikans[11]

Statistisk signifikans är ett begrepp som ofta missförstås. Nya och användarvänliga statistikprogram gör det lätt att göra beräkningar av olika slag, men en avigsida är att man tankemässigt kan distanseras från vad man faktiskt gör. En följd kan bli att man inte längre är

11 Punktestimering och konfidensintervall är ett vanligare och kanske mer intuitivt begripligt sätt att gestalta statistisk inferens, men min erfarenhet är att det är pedagogiskt ändamålsenligt att börja med hypotesprövning och signifikanstestning. Betydelsen av de antaganden som påverkar samplingfördelningens egenskaper kan vara tydligare vid hypotesprövning än vid estimering och konfidensintervall.

klar över att signifikans bygger på idén om hypotesprövning. Vid en statistisk hypotesprövning, i sin enklaste form, utgår man från antagandet att nollhypotesen (H_0) är sann, till exempel att medelvärdet i två populationer är lika ($\mu_I=\mu_C$). Alternativhypotesen (H_1) är ett antagande om att populationsmedelvärdena inte är lika ($\mu_I\neq\mu_C$).

Hypotesprövning och statistisk signifikans bygger på ett tankeexperiment. Anta att man drar två obundna och slumpmässiga urval (OSU) från en population. Därefter beräknas ett medelvärde för respektive grupp avseende variabeln av intresse, till exempel antal dagar i stabilt boende de senaste 6 månaderna (för f.d. hemlösa). Därefter beräknas skillnaden mellan medelvärdena. Skillnaden mellan dessa två medelvärden kommer sannolikt att ligga nära den sanna skillnaden, det vill säga skillnaden mellan populationsmedelvärdena ($\mu_I-\mu_C$).

Om det vore möjligt att upprepa ovanstående procedur ett stort antal gånger (med slumpmässiga urval och beräkningar av skillnader mellan urvalsmedelvärden), så skulle man få en fördelning där de flesta skillnaderna låg nära den ”sanna” skillnaden och färre skillnader en bit från den sanna skillnaden. Ju större antalet hypotetiska försök är, desto närmare kommer medelvärdet av medelvärdesskillnader ligga den sanna medelvärdesskillnaden, det vill säga medelvärdesskillnaden i populationen. När antalet försök går mot oändligheten får man en teoretisk fördelning som kallas samplingfördelningen (sampel=urval). Samplingfördelningen har ett eget medelvärde (av medelvärdesskillnader) samt en standardavvikelse vilken brukar kallas standardfelet (standard error).

Standardfelets storlek beror i sin tur på hur stora urvalen är samt hur stora de sanna standardavvikelserna (σ_I och σ_C) är i de två populationerna. Motsvarande standardfel för skillnaden mellan två medelvärden är $se(m_I-m_C)$, ekvation 9:

$$se\ (m_I - m_C) = \sqrt{\frac{\sigma_I^{\ 2}}{n_I} + \frac{\sigma_C^{\ 2}}{n_C}} \qquad (9)$$

Eftersom man sällan känner till de sanna standardavvikelserna i populationerna brukar standardfelet approximeras i enlighet med ekvation 10:

$$se\ (m_I - m_C) = \sqrt{\frac{(n_I - 1)sd_I^{\,2} + (n_C - 1)sd_C^{\,2}}{n_I + n_C - 2} \cdot \frac{n_I + n_C}{n_I \cdot n_C}} \qquad (10)$$

Vid signifikanstester av skillnaden mellan två medelvärden i två OSU är det följade hypotes man prövar (där μ_I och μ_C är populationsmedelvärden):

$$H_0: \mu_I - \mu_C = 0 \qquad (11)$$

Med en sann nollhypotes är det osannolikt att få en skillnad som ligger långt från den sanna skillnaden. Mer sannolikt är att den ligger nära den sanna skillnaden. Med stöd av samplingsfördelningen kan man beräkna hur sannolik en skillnad med en given storlek är givet en sann nollhypotes. Om den aktuella skillnaden är så stor att sannolikheten att få den är mindre än fem procent, brukar man säga att skillnaden är statistiskt signifikant och man förkastar H_0.

Genom att dividera den observerade skillnaden ($m_I - m_C$) med standardfelet får man ett Z-värde, ett värde i den standardiserade normalfördelningen[12]. Hypotesen förkastas vid ett dubbelsidigt test om …

$$Z_{\alpha/2} < \frac{m_I - m_C - 0}{se\,(m_I - m_C)} < -Z_{\alpha/2} \qquad (12)$$

… där $-Z_{\alpha/2}$ och $Z_{\alpha/2}$ anger gränser i den standardiserade normalfördelningen. Om Z-värdet ligger utanför dessa gränser, så anses skill-

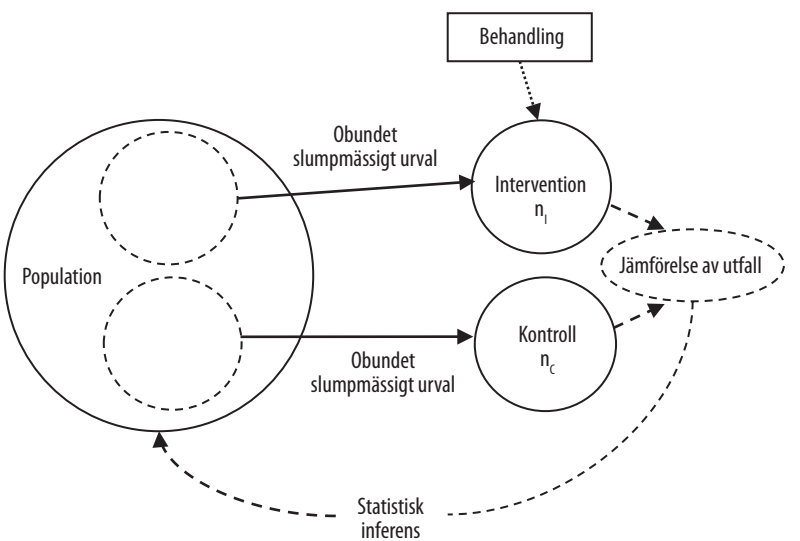
12 Det finns många andra fördelningar man kan använda för hypotesprövning och därmed för att räkna fram den statistiska signifikansen, till exempel *t*-fördelningen, $\chi 2$-fördelningen med mera beroende vilka statistiskt relevanta egenskaper utvärderingen har för ett specifikt utfallsmått.

naden vara statistiskt signifikant. Om signifikansnivån α är satt till 0,05, så är $-Z_{\alpha/2}$= -1,96 och $Z_{\alpha/2}$=+1,96.

Hela proceduren för en RCT illustreras i figur 16:3 (Byström, 1990, s. 205) såsom den är tänkt statistiskt att fungera. Man drar två *OSU* från en population. Deltagarna i den ena gruppen exponeras för interventionen medan den andra får ett kontrollalternativ. Därefter mäter man effekterna, räknar ut skillnaderna, prövar hypotesen statistiskt med hjälp ett signifikanstest för att slutligen dra slutsatser om populationen, alltså förkastar eller accepterar $H_{\circ}$. Att inte explicit specificera $H_{\circ}$, standardfel och beslutsregel medför, enligt min mening, en risk för att den statistiska signifikansen mystifieras så att tolkningen av resultaten försvåras.

För Cohens *d* beräknas standardfelet i enlighet med ekvation 13 (se Borenstein m.fl., 2009, s. 25–27).

$$se\,(d) = \sqrt{\frac{N}{n_I\, n_C} + \frac{d^2}{2N}} \qquad (13)$$



Figur 16:3. Det en RCT avser att approximera statistiskt.

Även när signifikans ska tolkas för *OR* och *RR* brukar varken H_o, standardfel eller beslutsregel formuleras i artiklar där resultat från *RCT* presenteras. För *OR* och *RR* brukar den outtalade H_o vara att det inte finns någon relativ skillnad mellan interventions- och kontrollgruppen där nollhypoteserna är följande (14 och 15):

$$H_0 : Odds_I / Odds_K = OR = 1{,}00 \qquad (14)$$

$$H_0 : Risk_I / Risk_K = RR = 1{,}00 \qquad (15)$$

H_o förkastas i båda fallen om effekten vägd mot standardfelet (enligt samma grundidé som för medelskillnader men med hjälp av exponentialfunktioner och logaritmer[13]) avviker tillräckligt mycket från det förväntade värdet 1,00 (givet en sann nollhypotes).

Att slumpmässigt fördela deltagarna i en undersökningsgrupp till en interventionsgrupp respektive en kontrollgrupp är inte samma sak som två slumpmässiga urval (jfr Shadish m.fl., 2002, s. 248). Man brukar dock använda formler som utvecklats för slumpmässiga urval då standardavvikelsen för samplingfördelningar för *RCT* ska beräknas.[14] Altman (1991, s. 86) framhåller följande:

> In a study with random allocation *the difference between treatment groups behave like the differences between random samples*. [vår kursivering]

Med en RCT försöker man, som antytts ovan, approximera den situation som illustreras i figur 16:3. Man drar två slumpmässiga urval. Individerna i det ena urvalet exponeras för interventionen medan individerna i det andra urvalet exponeras för kontrollalternativet.

13 För OR förkastas H_0 om $e^{Z_{\alpha/2} < \ln(OR) / \sqrt{1/A+1/B+1/C+1/D}}$ eller $e^{-Z_{\alpha/2} > \ln(OR) / \sqrt{1/A+1/B+1/C+1/D}}$ och för RR förkastas H_0 om $e^{Z_{\alpha/2} < \ln(RR) / \sqrt{(1/A-1/n_1)+(1/C+1/n_2)}}$ eller om $e^{-Z_{\alpha/2} > \ln(RR) / \sqrt{(1/A-1/n_1)+(1/C+1/n_2)}}$.

14 Samplingfördelningen för till exempel skillnaden mellan medelvärden D i två obundna och slumpmässigt dragna stickprov har ett standardfel som räknas fram på exakt samma sätt som i ekvation 10 ovan (Newbold, 1988, s. 365).

När rimlig tid förflutit räknar man fram skillnaden mellan medelvärdena för ett givet utfall i de båda grupperna samt testar om skillnaden är statistiskt signifikant och så vidare.

De slutsatser man drar från en signifikanstestning i en RCT är en form av generalisering. Om skillnaden inte är statistiskt signifikant, så betyder detta att den är så pass liten att den kan förklaras av skillnader som uppkommit av den slump som skapats vid randomiseringsprocessen givet en sann nollhypotes. Om däremot skillnaden är statistiskt signifikant, så är skillnaden för stor för att kunna förklaras av den slump som skapats genom randomiseringsprocessen givet en sann nollhypotes. Ett *p*-värde på mindre 0,05 innebär, som nämnts, att sannolikheten för den observerade skillnaden som uppkommit genom den slump som skapats genom randomiseringen är mindre än fem procent givet en sann nollhypotes.

Om man antar att en RCT verkligen är en lyckad approximation av två slumpmässiga urval som i figur 16:3, så kan man dra statistiska slutsatser om ett förväntat resultat i hela populationen med stöd av *p*-värdet. Om man inte antar detta, utan betraktar urvalet av dem som deltagit i undersökningen som icke-slumpmässigt, så är sådana slutsatser problematiska. Krause och Howard (2003, s. 762) tydliggör detta problem och lyfter fram vissa metodologiska svårigheter vid tolkandet av resultaten:

> Random assignment of patients to comparison groups creates the independent random sample quality of the comparison groups in a randomized trial. This is inductively meaningful insofar as the population that is represented by *those who are effectively randomly assigned is the population to which estimated treatment effects are expressly generalized*. When there is some slippage between intended population and actually represented population . . . it is hard to know what to make of pooled within-group variance as a basis for estimating the sampling error of the mean between group differences (i.e. to what patient population is this relevant). And so it is often hard to know what to make of the null hypothesis significance test in randomized clinical trials. [vår kursivering]

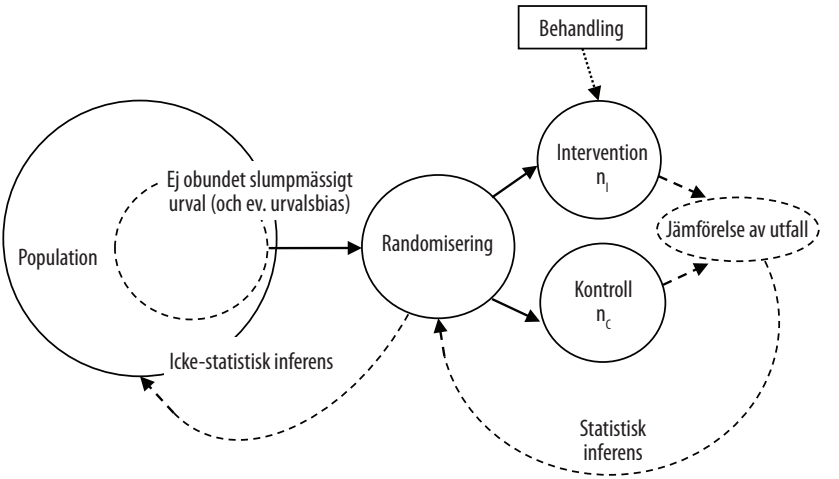
Att det kan finnas ett sådant glapp mellan dem som ingår i utvärderingen och den population man vill generalisera till lyfts fram av Rothwell (2005), till exempel med avseende på den kontext inom vilken utvärderingen genomförts, hur urvalet av patienter gjorts, de egenskaper som kännetecknar de utvalda patienterna, skillnader mellan de omständigheter som rådde vid utvärderingen och vardaglig praxis, utfallsmått och uppföljningstid samt mätning och rapportering av skadliga biverkningar.

Rothwell (2005, s. 84) illustrerar problemet i sin översikt med exempel från en utvärdering av en typ av kirurgiskt ingrepp. Översikten visar att fler faktorer påverkade resultaten än själva behandlingen. Ett exempel var klinikens snabbhet. Vid snabba kliniker hade interventionen bättre effekter än vid långsamma (allt annat lika). Om enbart långsamma kliniker skulle ha ingått i översikten hade det snedvridit resultaten. Att de utvalda patienterna randomiserats till interventions- respektive kontrollgrupp hjälper inte. Den osäkerhet som signifikanstestet kvantifierar med hjälp av *p*-värdet hanterar inte problemet i detta fall.

En alternativ tolkning av vad en RCT innebär illustreras i figur 16:4 (jfr Wright, 1997, s. 22; Lachin, 1988, s. 295[15]). Urvalet av den grupp som ingår i utvärderingen utgör inte ett OSU från den population man vill dra slutsatser om. Detta innebär i strikt mening att det inte finns någon kontrollerad slumpprocess med stöd av vilken man kan göra hypotesprövningar. Den statistiska inferensen, enligt detta sätt att se, räcker alltså inte längre än till den utvalda grupp som randomiserats. Det finns med andra ord ett urvalsbias där randomiseringen inte är till någon hjälp.

Ett sätt att hantera problemet med urvalsbias och extern validitet är att väga samman resultaten från flera olika RCT inom ramen

15 Lachin (1988) skriver om tre alternativa modeller: en ”sampling-based population model” (motsvarande figur 13:3), en ”invoked population model” (motsvarande figur 13.4) och en ”randomization model” (motsvarande figur 13.4, men endast den del som omfattas av den statistiska inferensen och alltså inte någon population).



Figur 16:4. Det en RCT kan innebära rent faktiskt.

för systematiska översikter i form av metaanalys. På detta sätt kan representativiteten förbättras något.

Det finns ytterligare problem som bör nämnas här som kan försvåra tolkningen av resultat från RCT. Ett problem utgörs av ett missförstånd rörande RCT och signifikans (se Altman, 1991, s. 170 & 489; Daniel, 1998, s. 27–28; Shadish m.fl., 2002, s. 42–43) är att tolka *p*-värdet som sannolikheten att ”resultatet uppkommit av slumpen”. Carver (1978, s. 383) har kallat missförståndet ”odds-against-chance fantasy” och han har bidragit med den kanske mest pedagogiska beskrivningen:

> [”Odds-against-chance” fantasy] is an interpretation of the *p* value as the probability that the research results were due to chance, or caused by chance /…/ As has been explained, the *p* value is the probability of getting the research results when it is first assumed that it is actually true that chance caused the results. It is therefore impossible for the *p* value to be the probability that chance caused the mean difference between two research groups since (a) the *p* value was calculated by assuming that the probability was 1.00 that chance did cause

> the mean difference, and (b) the *p* value is used to decide whether to accept or reject the idea that probability is 1.00 that chance caused the mean difference.

Ett exempel på ett liknande missförstånd angående statistisk signifikans har lyfts av Krause och Howard (2003, s. 761) – att signifikanstestning normalt sett är ett sätt att mäta om randomiseringen lyckats:

> It is sometimes suggested that the significance test of comparison-group outcome differences is a protection against the failure of random assignment to produce identical distributions of causal randomized-variable value combinations in each comparison group. In fact it is the other way around . . . random assignment protects significance testing. Without random assignment, a test of statistical significance of between-group differences is not conventionally interpretable.[16]

Detta tydliggör att begreppet statistisk signifikans är problematiskt utan en tydligt specificerad nollhypotes. Dessvärre är nollhypotesen ofta underförstådd och sällan explicit formulerad, vilket poängterats tidigare. Vidare kan inte sannolikheten för mer eller mindre stora observerade skillnader beräknas utan hänsyn till samplingfördelningen och standardfelet.

Sammanfattningsvis: varken nollhypotes, standardfel eller beslutsregel för förkastande av nollhypotesen brukar dessvärre formuleras explicit då forskningsresultat från *RCT* redovisas. Detta kan mystifiera tolkningen av statistisk signifikans. Ett kanske minst lika vanligt sätt att redovisa statiska resultat är dock att använda konfi-

16 Det kan nämnas att signifikanstester och konfidensintervall mycket ofta beräknas inom etablerad forskning trots problemet som lyfts fram i citatet av Krause och Howard. Vad man kan respektive inte kan göra har ju att göra med vilka antaganden man gör. Om man antar att urvalen i praktiken beter sig som om det rörde sig om slumpmässiga urval, så kan man naturligtvis både beräkna standardfel och tolka statistisk signifikans.

densintervall (vilket inte behandlas här av utrymmesskäl). Här utgår man inte från en nollhypotes, men problemet med mystifiering och tolkningsproblem kvarstår även om konfidensintervall kanske är intuitivt lättare att förstå än statistisk signifikans och *p*-värden. Det centrala i båda fallen är samplingfördelningen och standardfelet samt de antaganden man gör för att kunna beräkna dem.

Sammanfattning

Det är av central betydelse att dokumentera vad såväl interventionsgruppen som kontrollgruppen exponerats för samt att hålla isär effekt och komparativ effekt. De vanligast förekommande statistiska effektmåtten för kontinuerliga utfall är *D* och *d,* medan binära utfall oftast beräknas med hjälp av *OR*, *RR* och *RD*. Standardfelen approximeras vanligtvis vid RCT med hjälp av formler som tagits fram för OSU (sampling-based population model). Statistisk signifikans baseras på en nollhypotes med vars hjälp en samplingfördelning definieras och ett standardfel räknas fram. Ett alternativt och mer försiktigt sätt att se på tolkning av resultat från RCT är att tydliggöra problem med generalisering och extern validitet om de som deltar i utvärderingen inte utgör OSU:n från den population till vilken man vill generalisera. Av detta skäl är det viktigt att genomföra systematiska översikter och metaanalyser för att hantera problemet.

Fördjupningslitteratur

Altman, D. G. (1991). *Practical statisicts for medical research*. Boca Raton/ London/ New York/ Washington DC: Chapman & Hall/CRC.

Borenstein, M., Hedges, L., Higgins, J. P. T. & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester; West Sussex: John Wiley & Sons.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi-experimental design for generalized causal inference*. Boston & New York: Houghton Mifflin Company.

Referenser

Altman, D. G. (1991). *Practical statisicts for medical research*. Boca Raton/ London/ New York/ Washington DC: Chapman & Hall/CRC.

Borenstein, M., Hedges, L., Higgins, J. P. T. & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. Chichester; West Sussex: John Wiley & Sons.

Brozek, J. L., Akl, E. A., Alonso-Coello, P., Lang, D., Jaeschke, R., Williams, J. W., m.fl. (2009). Grading quality of evidence and strength of recommendations in clinical practice guidelines. *Allergy*, *64*, 669–677.

Byström, J. (1990). *Grundkurs i statistik*. Stockholm: Natur och Kultur.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*, 378–399.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. 2nd edition. Hillsdale New Jersey: Lawrence Erlbaum Associates, Publishers.

Daniel, L. G. (1998) Statistical significance testing: A historical overview of misuse and misinterpretations with implications for the editorial policies for the educational journals. *Research in the Schools*, *5*, 23–32.

Glenny, A. M., Altman, D. G., Song, F., Sakarovitch, C., Deeks, J. J., D'amico, R., m.fl. (2005). Indirect comparisons of competing interventions. Health Technology Assessment NHS R & D HTA Programme (www.hta.ac.uk).

GRADE Working Group (2004). Grading quality of evidence and strength of recommendations. *BMJ*, *328*, 1–8.

Haynes, R. B., Devereaux, P. J. & Guyatt, G. H. (2002). Clinical expertise in the era of evidencebased medicine and patient choice. *Evidence Based Medicine*, *7,* 36–8.

Hróbjartsson, A. & Gøtzsche, P. C. (2004). Placebo interventions for all clinical conditions. *The Cochrane Database of Systematic Reviews*, *Issue 2*.

Krause, M. S. & Howard, K. I. (2003). What random assignment does and does not do. *Journal of Clinical Psychology*, *49*, 751–766.

Lachin, J. M. (1988). Statistical properties of randomization in clinical trials. *Controlled Clinical Trials*, *9*, 289–311.

Luce, B. R., Kramer, J. M., Goodman, S. N., Connor, J. T., Tunis, S., Whicher, D., m.fl. (2009). Rethinking randomized controlled trials for comparative effectiveness research: The need for transformational change. *Annals for Internal Medicine 151,* 206–10.

Manchikanti, L., Falco, F. J. E., Boswell, M. V., Hirsch, J. A. (2010). Facts, fallacies, and politics of comparative effectiveness research: Part I. Basic considerations. *Pain Physician, 13*, E23–E54.

McHugo, G. J., Bebout, R. R., Harris, M., Cleghorn, S., Herring, O., Xie, H., m.fl. (2004). A randomized controlled trial of integrated versus parallel housing services for homeless adults with severe mental illness. *Schizophrenia Bulletin, 30*, 969–82.

Morse, G. A., Calsyn, R. J., an Klinkenberg, W., Helminiak, T. W., Wolff, N., Drake, R. E., m.fl. (2006). Treating homeless clients with severe mental illness

and substance use disorders: costs and outcomes. *Community Mental Health Journal, 42,* 377–404.
Newbold, P. (1988). *Statistics for business and administration*. Englewood Cliffs, New Jersey: Prentice-Hall international Editions.
Piaggio, G., Elbourne, D. E., Altman, D. G., Pocock, S. J. & Evans S. J. W., for the CONSORT group (2006). Reporting of noninferiority and equivalence randomized trials an extension of the CONSORT statement. *JAMA, 295*, 1152–1160.
Rothwell, P. M. (2005). External validity of randomized controlled trials: "to whom do the results of this trail apply?" *Lancet, 365*, 82–93.
Shadish, W. R., Cook, T. D. & Campbell, D. T (2002). *Experimental and quasi-experimental design for generalized causal inference*. Boston & New York: Houghton Mifflin Company.
Wright, D. B. (1997). *Understanding statistics. An introduction for the social sciences*. London/ Thousand Oaks/ New Delhi: SAGE Publications.

Kapitel 17 | Lars-Göran Öst

Klinisk signifikans

Ända sedan 1970-talet har man talat om vikten av att komplettera testningarna av den statistiska signifikansen i en studie med analys av den kliniska signifikansen (Kazdin, 1977; Lick, 1973; Wolf, 1978). Kortfattat kan man säga att klinisk signifikans handlar om det praktiska värdet eller betydelsen av behandlingseffekten, det vill säga om den gör någon verklig skillnad för klienten eller för anhöriga. För att illustrera frågan kan vi anta att en forskare undersöker en behandlingsmetod för hypertoni och att behandlingsgruppens medelvärde för blodtryck i vila före behandlingen var 150 mmHg. Efter behandlingen hade medelvärdet sjunkit till 148 mmHg. Denna skillnad är statistiskt signifikant om man har ett tillräckligt stort urval men den är inte kliniskt signifikant eftersom patienterna fortfarande befinner sig i riskzonen för olika kardiovaskulära sjukdomar.

Tre sätt att utvärdera klinisk signifikans har utvecklats (Kazdin, 2003): (1) Subjektiv evaluering, (2) Social inverkan och (3) Jämförelsemetoder.

Subjektiv evaluering

Användning av subjektiv evaluering innebär att man bestämmer betydelsen av patientens beteendeförändring genom att mäta åsikterna hos de individer som har kontakt med honom eller henne,

till exempel anhöriga eller vänner. Andras åsikter är relevanta eftersom de ofta har en kritisk roll i att identifiera, definiera och interagera med personer som de anser vara dysfunktionella eller avvikande. Om behandlingen fungerar och har en viktig inverkan borde den leda till märkbara skillnader för patienterna själva och personer i deras närmiljö.

Problem med subjektiv evaluering

Globala skattningsskalor används vanligen för att mäta subjektiva evalueringar och dessa har inte så hög stabilitet över tid (test–retest-reliabilitet). Ofta kan det vara så att en patient som har en komplex problematik blir skattad vad gäller det svåraste problemet före behandlingen. Om behandlingen är framgångsrik kanske den anhöriga tenderar att ”glömma bort” hur stora svårigheter patienten hade i det avseendet och i stället skattar ett annat problem som nu kommit i fokus. Därigenom blir förbättringen mindre än om man bedömt hela patientens problematik. Det kan också vara så att patienten eller personer i hans eller hennes närhet märker skillnader i beteendet som en funktion av terapin, men det behöver inte betyda att patienten har förändrats särskilt mycket. Förväntan om att behandlingen ska ha effekt kan leda till överskattningar. Över huvud taget måste man vara försiktig med subjektiva evalueringar då det är möjligt att dessa kan visa förändringar när andra mått inte gör det.

Social inverkan

Ett annat sätt att evaluera klinisk signifikans handlar om utfall som mäts i vardagslivet och som är viktiga för samhället i stort. Exempel på detta är frekvens av brott, skolskolk, rattfylleri, sjukdomar, sjukhusvistelser och självmord. Mått på social inverkan har ofta använts i kliniska och tillämpade studier, till exempel i preventionsprogram med fokus på barn som är i riskzonen för att utveckla psykiska eller kroppsliga problem.

Problem med social inverkan

Det finns en mängd problem med social inverkan. För det första är det fråga om grova mått som kan utsättas för en mängd andra influenser än av själva behandlingen. Dessutom är det ofta så att dessa mått inte registreras särskilt reliabelt och frekvensen fel i måtten kan vara relativt hög. Vidare är instrumenteffekter (se kapitel 4) ofta ett problem för mått på social inverkan. Slutligen tar allmänheten, media och politiker ofta fasta på mått på social inverkan som om de utgör kärnpunkten för att evaluera värdet och effektiviteten hos ett behandlingsprogram. Mått på social inverkan måste evalueras och tolkas på ett genomtänkt sätt. Dilemmat med dessa mått kommer från deras mest framträdande karakteristika; den höga trovärdigheten och de oftast dåliga psykometriska egenskaperna hos måtten.

Jämförelsemetoder

Vid behandlingens slut kan patienten jämföras med någon standard för att bestämma om förändringen är kliniskt signifikant. Man kan använda två typer av jämförelser:

- Normativa jämförelser innebär att man jämför personens prestation med andra personers.
- Ipsativa jämförelser innebär att jämföra individen med sig själv.

Normativa sampel

Om man har tillgång till data från en normalgrupp på det aktuella måttet så är det naturligt att fråga i vilken utsträckning patienterna faller inom den normativa fördelningen efter att ha genomfört behandlingen. Man räknar ut hur stor andel av patienterna i respektive behandlingsgrupp som ligger innanför den normativa fördelningen och testar om grupperna skiljer sig åt i detta avseende.

Dysfunktionella sampel

Om det inte finns normativa data på måttet kan man vid behandlingens slut jämföra patienternas värden med värden från obehandlade

patienter och de bör då avvika markant. Ett ofta använt kriterium är två standardavvikelser (*SD*) från medelvärdet för patientsamplet. Ligger patientens värde utanför denna gräns (i funktionell riktning) så anses personen ha en kliniskt signifikant förändring.

Problem med jämförelsemetoder

Det finns ett antal frågor som aktualiseras när man använder sig av jämförelsemetoder. En viktig fråga är vilka personer som ska ingå i den normativa gruppen. Ska det vara ett slumpmässigt urval ur befolkningen och då inkludera en andel som har eller har haft psykiska störningar (46 procent i en relativt färsk amerikansk epidemiologisk studie av Kessler, Berglund, Demler, Jin, Merikangas & Walters, 2005)? Eller ska personerna som slumpats fram genomgå en diagnostisk intervju för att kunna exkludera dem som har eller har haft en psykisk störning? Då får man en supernormal grupp som har ännu lägre värden på det aktuella måttet.

Andra frågor som aktualiseras är att även om en normativ grupp kan identifieras – exakt vilka beteenden ska definieras inom den normativa nivån? Vi vet att symtom är vanliga hos normativa sampel. Livskvalitet är lika viktigt eller till och med viktigare för att definiera klinisk signifikans.

Beräkning av kliniskt signifikant förbättring

Jacobson, Follette och Revenstorf (1984) och Jacobson och Truax (1991) har beskrivit statistiska metoder för att beräkna kliniskt signifikant förändring. För att den förbättring som patienten uppvisar efter behandlingen ska anses kliniskt signifikant måste följande två kriterier vara uppfyllda:

1. Reliable Change Index (RCI)

Förändringen som den enskilde patienten har fått ska vara tillräckligt stor för att vara *statistiskt reliabel* och inte slumpmässig ($p<0.05$). Formeln för att beräkna detta är:

Differensen (före–efter)/$\sqrt{2}(S_E)^2$ ska vara ≥ 1.96

där $S_E = SD\sqrt{(1 - r)}$ och *r* är måttets interna konsistens (Cronbachs α). Jacobson och Truax (1991) använde måttets test–retest-reliabilitet men Lambert, Hansen och Bauer (2008) citerar flera studier som rekommenderar att den interna konsistensen används.

2. Gränsvärde

Det finns möjlighet att beräkna tre olika gränsvärden. Patientens värde ska ligga

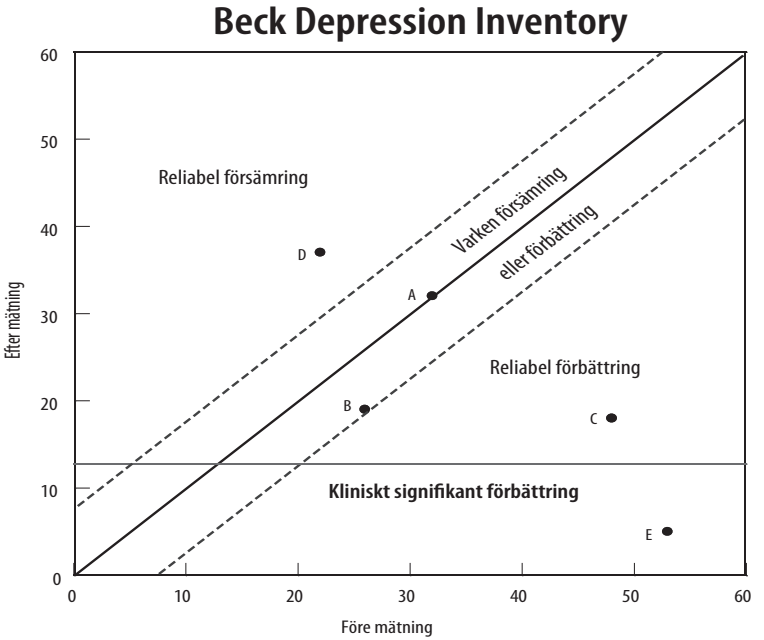
a) innanför normalgruppens variationsområde definierat som:
$M_{normalgrupp} + 2SD_{normalgrupp}$

b) utanför patientgruppens variationsområde (före behandling) definierat som:
$M_{patienter} - 2SD_{patienter}$

c) innanför den vägda mittpunkten mellan normalgruppens och patientgruppens medelvärden definierat som:
$((SD_{patienter} \times M_{normalgrupp}) + (SD_{normalgrupp} \times M_{patienter}))/(SD_{patienter} + SD_{normalgrupp})$

Gränsvärde C rekommenderas när det är en överlappning mellan fördelningarna för normalgrupp och patienter. När det inte är en överlappning mellan fördelningarna rekommenderas gränsvärde A och när det saknas data från en normalgrupp är gränsvärde B det enda som är möjligt. Det kan också påpekas att patientens förevärde måste ligga utanför normalgruppens variationsområde om det ska vara meningsfullt att inkludera denna patient i beräkningen av andelen som uppfyller gränsvärdet.

I figur 17:1 illustreras RCI och gränsvärdet för Beck Depression Inventory baserat på data från Ogles, Lambert och Masters (1996).



Figur 17:1. Illustration av de fyra utfallskategorierna med hjälp av BDI.

Fyra utfallskategorier

Genom att använda de två kriterier som Jacobson och Truax (1991) beskrivit kan patienterna klassificeras i fyra utfallskategorier:

- Försämrade: RCI visar signifikant försämring
- Oförändrade: RCI visar varken signifikant försämring eller signifikant förbättring
- Förbättrade: RCI visar signifikant förbättring
- Återställda: RCI visar signifikant förbättring och patientens poäng uppfyller kriteriet för gränsvärdet (under det vågräta strecket i figur 17:1).

Tabell 17:1 illustrerar användningen av dessa utfallskategorier i en studie av Okiishi, Lambert, Eggett, Nielsen och Dayton (2006) med 71 terapeuter och 6499 patienter (collegestudenter). Med hjälp av

Tabell 17:1. Utfallet (i procent) för de bästa och de sämsta terapeuterna i Okiishis m.fl. (2006) studie med tillämpning av Jacobsons fyra utfallskategorier.

	Utfallskategori			
Terapeuter	**Försämrad**	**Oförändrad**	**Förbättrad**	**Återställd**
Bästa	5,2	50,9	21,5	22,4
Sämsta	10,6	61,4	17,4	10,6

Outcome Questionnaire-45 (Lambert, Burlingame, Umphress, Hansen, Vermeersch, Clouse, Glenn m.fl, 1999) rangordnades terapeuterna på basis av det genomsnittliga resultat som deras patienter fått. RCI var 14 poäng och gränsvärdet 63 poäng i denna studie. Man jämförde sedan de 10 procent bästa med de 10 procent sämsta terapeuterna och fann en signifikant skillnad i utfall.

Något som är anmärkningsvärt i denna studie är att även bland de patienter som hade turen att hamna hos de bästa terapeuterna så är det mer än hälften (56 procent) som inte fick ett positivt utfall av terapin. Om detta beror på att terapeuterna inte är särskilt duktiga, patienterna särskilt svåra eller att OQ-45 inte är så känsligt för förändring kan inte bedömas från studien.

Betydligt bättre resultat erhölls av 294 psykologistudenter vid Stockholms universitet som under handledning behandlade 591 patienter med kognitiv beteendeterapi (Öst, Karlstedt & Widén, 2012). Studenterna gick termin 7–9 av den femåriga psykologutbildningen och patienterna hade framför allt ångeststörningar eller depression med en genomsnittlig varaktighet av cirka 15 år. Resultaten framgår av tabell 15:2.

Tabell 17:2. Utfallet (i procent) för psykologistudenter på Beck Anxiety Inventory och Beck Depression Inventory.

	Utfallskategori			
Mått	**Försämrad**	**Oförändrad**	**Förbättrad**	**Återställd**
Beck Anxiety Inventory	1,1	17,1	19,7	63,2
Beck Depression Inventory	1,3	27,7	11,0	60,0

Begränsningar vad gäller klinisk signifikans

Lambert med flera (2008) beskriver ett antal studier som har tillämpat andra formler än Jacobson och Truax men finner att överlappningen mellan resultaten för de olika formlerna är stor. De rekommenderar att man ska använda Jacobsons och Truax metodik (det vill säga RCI och gränsvärde) som rutinmetod i effektutvärderingar.

De tar också upp några begränsningar med den nuvarande metodiken. Ett problem är patienter som redan när de påbörjar behandlingen ligger innanför normalgruppens fördelning som den definierats ovan. Det innebär att dessa personer inte kan uppfylla gränsvärdeskriteriet även om de kan uppnå RCI. För denna subgrupp av patienter kan man begränsa sig till RCI och kalla det för reliabel förbättring om de uppfyller kriteriet. Ett annat problem finns med populationer av kroniska patienter där kliniskt signifikant förbättring kan vara ett opraktiskt mål. Ett förslag som tas upp av Lambert med flera (2008) är att inom patientgruppen definiera en dysfunktionell fördelning, bestående av inneliggande patienter, och om patienten uppfyller ett gränsvärde definierat av öppenvårdspatienter skulle kriteriet vara uppfyllt. Ett ytterligare problem som författarna tar upp är att det saknas normdata för många av de mått som används i psykoterapiforskning. Detta kommer dock att förbättras med ny forskning.

Inte längre uppfylla diagnoskriterierna för störningen

Ett alternativt mått på klinisk signifikans som ibland förekommer är att efter behandlingen genomföra en ny diagnostisk intervju för att kunna bedöma om patientens diagnostiska status har förändrats. Det ligger något tilltalande i att individen inte längre uppfyller diagnoskriterierna; det antyder att störningen har försvunnit eller botats. Men att inte längre uppfylla diagnoskriterierna behöver inte innebära någon markant förbättring. Ta egentlig depression som exempel. Symtomkriteriet listar åtta olika symtom och patienten måste ha minst fem för att få diagnosen. Anta att man efter be-

handlingen har fyra symtom, det vill säga en liten förändring, men då uppfyller man inte längre diagnosen egentlig depression. Detta sätt att evaluera klinisk signifikans beskriver en mycket heterogen grupp patienter; från dem som helt har blivit av med störningen till sådana som har en minimal förbättring.

Fördjupningslitteratur

Jacobson, N. S. & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12–19.

Kazdin, A. E. (1977). Assessing the clinical and applied importance of behavior change through social validation. *Behavior Modification, 1,* 427–452.

Kazdin, A. E. (2003). *Research design in clinical psychology*, 4th ed. Boston: Allyn & Bacon.

Referenser

Jacobson, N. S., Follette, W. C., Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15*, 336–352.

Jacobson, N. S. & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12–19.

Kazdin, A. E. (1977). Assessing the clinical and applied importance of behavior change through social validation. *Behavior Modification, 1,* 427–452.

Kazdin, A. E. (2003). *Research design in clinical psychology*, 4th ed. Boston: Allyn & Bacon.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R. Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry, 62*, 593–602.

Lambert, M. J., Burlingame, G. M., Umphress, V., Hansen, N. B., Vermeersch, D. A., Clouse, Glenn C., m.fl., (1999). The reliability and validity of the Outcome Questionnaire. *Clinical Psychology & Psychotherapy, 3*, 249–258.

Lambert, M. J., Hansen, N.B. & Bauer, S. (2008). Assessing the clinical significance of outcome results. I: A.M. Nezu & C.M. Nezu (Red.) *Evidence-based outcome research*. Oxford: Oxford University Press.

Lick, J. (1973). Statistical vs. clinical significance in research on the outcome of psychotherapy. *International Journal of Mental Health, 2,* 26–37.

Ogles, B. M., Lambert, M. J. & Masters, K. S. (1996). *Assessing outcome in clinical practice*. Boston: Allyn & Bacon.

Okiishi, J. C., Lambert, M. J., Eggett, D., Nielsen, L. & Dayton, D. D. (2006). An analysis of therapist treatment effects: Toward providing feedback to individual therapists on their clients' psychotherapy outcome. *Journal of Clinical Psychology, 62*, 115–1172.

Wolf, M. M. (1978). Social validity: The case for subjective measurement of how applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis, 11*, 203–214.

Öst, L-G., Karlstedt, A. & Widén, S. (2011). The effects of CBT delivered by students in a psychologist training program. An effectiveness study. *Behavior Therapy, 43*, 160–173.

Kapitel 18 | Terje Ogden

Rapportering av randomiserat kontrollerade utvärderingar[1]

Randomiserade kontrollerade utvärderingar (eng. *Randomized Controlled Trials*, RCT) lämpar sig bättre än andra designer när det gäller att testa kausala samband mellan intervention och utfall eftersom den slumpmässiga fördelningen till experiment och kontrollgrupp innebär att grupperna i teorin blir ekvivalenta. RCT benämns inom interventionsforskning ofta som ”The Golden standard” (Shadish, Cook & Campbell, 2002).

I forskningslitteraturen finns många beskrivningar av hur en randomiserad kontrollerad studie bör designas, genomföras och rapporteras. En av de mest inflytelserika publikationerna för hur RCT-studiet ska rapporteras är *The CONSORT (Consolidated Standards of Reporting Trials) statement* (Schulz, Altman, Moher for the CONSORT group, 2010; Moher, Hopewell, Schulz, Montori, Gøtzsche, Devereaux m.fl., 2010)[2]. CONSORT:s riktlinjer har utarbetats av en grupp forskare och redaktörer från olika vetenskapliga

1 Kapitlet har översatts från norska till svenska av Catrine Kaunitz, Socialstyrelsen.
2 Utöver CONSORT finns riktlinjer för hur icke-randomiserade utvärderingar ska rapporteras, *The Transparent Reporting of Evaluations with Non-randomized Designs* (TREND) (Des Jarlais, Lyles, Crepaz & TREND Group, 2004) samt hur observationsstudier inom epidemiologi ska rapporteras, *The Strengthening the Reporting of Observational Studies in Epidemiology* (STROBE) (von Elm, Altman, Egger, Pocock, Gøtzsche, Vandenbroucke for the STROBE Initiative, 2008).

tidskrifter. En del av kriterierna präglas av att de primärt riktar sig till utvärderingar av medicinska interventioner, men rekommendationerna har fått en bredare användning och tillämpas i dag även inom discipliner såsom psykologi och socialt arbete. Vissa av rekommendationerna är dock inte direkt tillämpbara på psykosociala interventioner, vilket gör att smärre modifieringar och tillägg kan vara nödvändiga.

De första riktlinjerna från CONSORT publicerades 1996 och dessa har sedan uppdaterats 2001 och 2010 (www.consort-statement.org). Riktlinjerna sammanfattas i form av en checklista (tabell 1) och ett flödesschema (figur 1). Listan har följande huvudrubriker: introduktion, metod, resultat och diskussion. Redogörelsen för CONSORT:s riktlinjer i det här kapitlet har samma disposition som checklistan. Kapitlet innehåller också kommentarer från annan relevant litteratur. Fokus i sammanfattningen är den vanligaste typen av randomiserade studier, nämligen den där två grupper jämförs (*two group, parallell designs*).

Tidskrifter kräver ofta att CONSORT:s riktlinjer kombineras med andra för tidskriften specifika krav om hur ett artikelmanuskript ska vara utformat för att bli aktuellt för publicering. Ofta hänvisas också till manualen (2006) American Psychological Association (APA) som innehåller etablerade kriterier för hur vetenskapliga artiklar bör utformas.

Tabell 18:1. Rapportering av randomiserat kontrollerade utvärderingar (jfr Moher m.fl. 2001; 2010).

Titel och sammanfattning	Av titeln ska det framgå att det är en *randomiserad utvärdering* (1a). En strukturerad sammanfattning av utvärderingens design, metod, resultat och slutsatser (1b).
INTRODUKTION	
Bakgrund och syfte	Vetenskaplig bakgrund och syfte (2a). Specifika mål och hypoteser (2b).
METOD	
Design	Beskrivning av design, inklusive fördelningen (ratio) mellan experiment- och kontrollgrupp (3a). Viktiga ändringar i metod efter utvärderingens start, inklusive motiv till dem (3b).
Deltagare	Inklusionskriterier för medverkan (4a). Platser där data samlats in (4b).
Interventioner	En beskrivning av interventionen för respektive grupp så att de kan replikeras samt en beskrivning av hur och när de administrerades (5).
Utfallsmått	Klart definierade primära och sekundära effektmått, inklusive hur och vilka som utförde mätningarna (6a). Eventuella förändringar i val av utfallsmått och orsaken till att det har gjorts (6b).
Urvalsstorlek	Hur urvalsstorleken bestämdes (7a). Om aktuellt, förklaring till preliminära analyser och motiv till att fortsatt inkludering avbrutits (7b).
Randomisering/sekvensering	Metod för att generera slumpmässig fördelningssekvens (8a). Typ av randomisering och eventuella restriktioner (t.ex. blockförsök, stratifiering) (8b).
Randomisering/fördelningsmetod	Metod för att implementera den slumpmässiga fördelningssekvensen (t.ex. numrerade kapslar) och om sekvensen dolts för dem som berördes (9).
Randomisering/implementering	Vem som ansvarade för randomiseringssekvensen samt registrerade och fördelade deltagare till interventionerna (10).
Maskering	Om det förekom, för vem och hur grupptillhörighet maskerats (t.ex. deltagare, behandlare, datainsamlare) (11a). Om relevant, beskrivning av likheter mellan interventioner (11b).
Statistiska metoder	Statistiska metoder för att jämföra grupperna i primära och sekundära effektmått (12a). Metoder for kompletterande analyser (t.ex. subgruppsanalys, justerad analys) (12b).

RESULTAT	
Bortfall av deltagare i varje fas (diagram rekommenderas starkt).	För varje grupp, hur många som randomiserats till deltagande, som fått behandlingen, som fullföljt behandlingen och som ingår i analyserna (13a). Orsaker till bortfall för varje grupp (13b).
Rekrytering	Datum för rekryterings- och uppföljningsperioden (14a). Varför studien avslutats eller avbrutits (14b).
Baslinjedata	En tabell som visar demografiska och kliniska kännetecken för varje grupp (15).
Antal analyserade	Antal deltagare i varje grupp för respektive analys samt om analyserna utförts på de ursprungliga randomiserade grupperna (16).
Utfall och estimat	Resultat för varje primär och sekundär resultatvariabel och separat för varje grupp, inklusive den estimerade effektstorleken och dess precision (t.ex. 95 procent konfidensintervall) (17a). För binära effektmått, ange både absoluta och relativa effektstorlekar (17b).
Tilläggsanalyser	Andra analyser som genomförts, inklusive subgruppsanalyser och justerade analyser, med uppgift om vilka som är specificerade i förhand respektive explorativa (18).
Skador	Rapport av alla skador eller oförutsedda effekter i varje interventionsgrupp (19).
DISKUSSION	
Begränsningar	Undersökningens begränsningar, orsaker till potentiella snedvridningar av resultat (bias), bristande precision samt i förekommande fall massignifikansproblem (20).
Generaliserbarhet	Resultatens generaliserbarhet (extern validitet, användbarhet) (21).
Tolkning	Tolkning i enlighet med resultaten. Balans mellan nytta och skada och resultatens förhållande till annan relevant evidens (22).
ANNAN INFORMATION	
Registrering	Registreringsnummer och namn på registret (23).
Protokoll	Om aktuellt, var forskningsprotokollet finns tillgängligt (24).
Finansiering	Finansiärer och annat stöd, finansiärernas roll (25).

Titel och sammanfattning

Titel (1a)

För att underlätta sökningar i elektroniska databaser bör det framgå i titeln att det rör sig om en randomiserad studie.

Strukturerad sammanfattning (1b)

Det är viktigt att sammanfattningen är tydlig och detaljerad eftersom många endast läser sammanfattningen och avgör utifrån den om de ska läsa resten av artikeln. Sammanfattningen ska inte innehålla information som inte också finns i huvudtexten. Sammanfattningen bör beskriva design, metod, resultat och konklusion, i den ordningen.

Introduktion

Bakgrund och syfte (2a)

Normalt sett beskrivs bakgrunden till studien i löpande text. Vidare bör det finnas en generell beskrivning av undersökningens syfte och den kunskapslucka studien avser att fylla. Inledningen kan också innehålla en beskrivning av fördelar och möjliga nackdelar med interventionen samt en genomgång av eventuella kunskapsöversikter när det gäller metoden.

Frågeställningar och hypoteser (2b)

De frågeställningar som studien avser att söka svar på ska beskrivas. Det rör sig oftast om huruvida en specifik metod är effektiv eller inte. Hypoteser är mer specifika än frågeställningar och preciserar vilka problemställningar som ska testas med statistiska analyser för att besvara frågeställningarna.

Metod

Design

Design (3a)

Designen beskriver hur en undersökning är organiserad men ska också innehålla mer specifika delar av processen som exempelvis vilken metod som använts för att slumpa individer mellan de olika betingelserna och den relativa fördelningen mellan grupperna (eng. *allocation ratio*).

Eventuella förändringar i design efter att undersökningen påbörjats (3b)

Även om de flesta studier har en detaljerad plan för hur en undersökning ska gå till så inträffar ofta oförutsedda händelser som gör att planeringen måste ändras. Ändringar i design kan till exempel bero på att det uppstår problem med finansieringen eller att rekryteringen av deltagare till studien går sämre än förväntat. Sådana omständigheter kan leda till ändringar i inklusionskriterierna, tidpunkt för uppföljningar eller annat som att man utesluter en av undersökningsplatserna eftersom kvaliteten på data är för dålig. Oavsett orsak ska alla ändringar rapporteras.

Deltagare

Inklusionskriterier för deltagare (4a)

Beskrivning av inklusionskriterierna är avgörande för hur resultaten ska tolkas och för vilken population som resultaten är giltiga och tillämpbara (d.v.s. extern validitet). Huruvida deltagarna självrekryterats eller remitterats är ett exempel på något som kan ha betydelse vid tolkningen av resultat. Exklusionskriterier beskriver vad som kännetecknar personer som interventionen inte är lämpliga för. Den vanliga åtskillnaden mellan inklusions- och exklusionskriterier är enligt CONSORT onödig då samma kriterier kan användas både för att exkludera och inkludera deltagare.

Miljöer där data samlas in (4b)

Eftersom remitterande instanser och olika verksamheter där interventionerna prövas kan skilja sig åt med hänsyn till organisering, erfarenheter och resurser är det nödvändigt att miljöerna (eng. *settings*) för datainsamling beskrivs. Beskrivningen bör innehålla information där det framgår vad som utmärker miljön som studien genomförts i, inklusive information om sociala, ekonomiska och kulturella faktorer. Beskrivningen ska bland annat klargöra om det rör sig om storstad eller landsbygd och om det var privat verksamhet eller offentlig. Beskrivningen ska vara så grundlig att läsaren ska kunna bedöma om resultaten är tillämpbara för den egna verksamheten. Rapporteringen ska också innehålla information om på vilket sätt den studerade verksamheten avviker från reguljär verksamhet. Logistiska och praktiska problem när det gäller genomförandet av studien bör också rapporteras.

Interventioner (5)

Informationen om genomförande och administration av de studerade interventionerna ska vara så detaljerad att det är möjligt att replikera studien. Rapporteringen ska vara så grundlig att det ska vara möjligt för en professionell att börja arbeta enligt metoden. Om kontrollgruppen får standardbehandling (eng. *treatment as usual*) måste dessa insatser beskrivas lika grundligt som experimentgruppens behandling.

Det är den relativa effekten av interventioner som undersöks i randomiserade studier; hur stor är förändringen i experimentgruppen i jämförelse med kontrollgruppen? De vanligaste kontrollbetingelserna är standardbehandling, ingen intervention, väntelista för behandlingsinterventionen eller en mindre omfattande eller annan intervention. Jämförelser med kontrollgrupper som inte får några insatser alls eller som står på väntelista ger i regel i starkare effekter. Nedan följer en genomgång av olika jämförelsegrupper.

Jämförelsegrupp som inte får någon behandling alls

Fördelen med en design där jämförelsegruppen inte får någon intervention är att resultaten beskriver teoretiskt ”rena” effekter som underlättar jämförelser mellan olika interventioner som riktar sig till samma tillstånd (t.ex. aggressivitet). Ett problem med denna design är att personer som inte erbjuds någon behandling kan söka hjälp på annat håll. Det kan också finnas etiska invändningar mot att inte erbjuda behandling till personer som söker hjälp för potentiellt allvarliga problem.

Ett alternativ är att använda väntelista som jämförelsegrupp. Den gruppen får behandling vid ett senare tillfälle, i regel direkt efter eftermätningen eller uppföljningsmätningen. Det finns dock kritik mot att använda väntelista. Om det finns signifikanta skillnader mellan en aktiv intervention och en väntelistekontroll är den enda slutsats som kan dras enligt Chambless och Hollon (1998) att det är bättre att göra någonting än att inte göra någonting alls. En annan begränsning med väntelista är att det inte går att studera långtidseffekter i förhållande till kontrollgruppen. Replikationer av studien med resultat i samma riktning kan emellertid stärka tilltron till resultaten. Om behandlingen ger positiva resultat, oavsett i vilken omfattning, och om dessa resultat återkommer i flera studier så är behandlingen förmodligen kliniskt värdefull och ”nyttig”.

Jämförelsegrupp med andra behandlingar

Jämförelser med andra aktiva interventioner innebär att man kan kontrollera processer som är gemensamma för alla former av behandling och som är oberoende av vilken sorts behandling det rör sig om. Exempel på det är ”uppmärksamhetseffekter” från en intresserad behandlare eller forskare eller att klienten känner press att ändra sig (Chambless & Hollon, 1998). Om behandlings- och kontrollgruppen får behandlingar som liknar varandra så reduceras skillnaden i effekter (Kazdin, 2008). En sådan situation kan uppstå om man jämför två behandlingar som båda har effekter eller om ordinarie verksamhet håller hög kvalitet. Jämförelser med standardbe-

handling kontrollerar emellertid inte för ”kompensatorisk rivalitet”, något som kan förekomma när behandlarna i jämförelsegruppen anstränger sig för att få bättre resultat än de vanligtvis gör eftersom de ingår i utvärderingen som kontrollgrupp (Cunningham, 2002). Ett annat problem är att mekanismerna i experimentgruppens intervention ”läcker” över till den reguljära verksamheten, vilket kan reducera skillnaden mellan experimentgrupp och jämförelsegrupp. Därför måste kontrollgruppens behandling beskrivas utförligt. Samtliga interventioner bör beskrivas med avseende på innehåll, dos och kvalitet på utförandet.

Effektmått

Utfallsmått (6a)

De flesta studier använder flera parallella utfallsmått. Några av dessa är primära och andra är sekundära. Det primära utfallsmåttet är det som används i beräkningar av statistisk styrka och som antas ha störst betydelse för intressenter. CONSORT avråder från att använda flera primära utfallsmått eftersom resultaten då kan bli svåra att tolka. Denna rekommendation följs dock sällan i pedagogiska och psykosociala effektutvärderingar.

Om forskare definierar och rapporterar primära och sekundära utfallsmått kan samma användas av andra och resultaten kan jämföras och läggas samman med andra studier. Etablerade och validerade skalor eller konsensusbaserade mått bör användas. För att säkra kvaliteten på mätningarna ska skalornas psykometriska egenskaper alltid rapporteras. CONSORT avråder från att använda opublicerade mätinstrument eftersom det kan leda till systematiska fel i resultaten.

Om resultaten ska mätas vid flera tillfällen så bör forskarna på förhand bestämma vilken tidpunkt som ska tillmätas störst betydelse. Hur aktuella eller relevanta olika utfallsmått är beror på studiens syfte: utfallen bör vara anpassade till interventionen och kunna fånga upp omedelbara förändringar hos deltagarna efter avslutad behandling, till exempel minskat utagerande beteende. Samtidigt

kan det vara aktuellt att komplettera med längre uppföljningstider. I behandling av till exempel allvarliga beteendeproblem kan de primära utfallen vara knutna till aggressivitet och utagerande beteende medan sekundära utfall kan vara relaterade till internaliserade problembeteenden. Idealt bör man i interventionsstudier använda sig av *multimetoder* (flera sätt att mäta samma sak) och *multiinformanter* (flera uppgiftslämnare). Att endast använda sig av en grupp av uppgiftslämnare rekommenderas inte (Chambless & Hollon, 1998). Exempel på multimetoder är observation, test och intervjuer medan multiinformanter kan vara föräldrar, lärare, elever samt oberoende observatörer. Multiinformant- och multimetodstudier är också som regel ”multi-setting”, det vill säga genomförs på flera ställen, exempelvis hemma och i skolan. Genom att täcka in flera arenor ges möjlighet till jämförelser och generaliseringar mellan skola och hemma och omvänt.

Vissa utfallsmått kräver inte reliabilitetstest då de har hög ”face validity”, som exempelvis antalet arresteringar eller placeringar utanför hemmet. Övriga mått i randomiserade studier bör vara reliabla och valida i tidigare forskning (APA, 2001). Instrumentets begreppsvaliditet bör också värderas, det vill säga om instrumentet mäter det som avses att mäta. En viktig och problematisk fråga är mätinstrumentens sensitivitet för förändring. Instrument kan vara sensitiva när det gäller individuell förändring vid en given tidpunkt men inte nödvändigtvis lika bra på att fånga förändring över tid. Mätinstrument med hög sensitivitet för förändring innebär goda möjligheter att hitta signifikanta behandlingseffekter på kort sikt medan det kan vara osäkert huruvida de fångar förändring på lång sikt. Bristen på forskning är stor när det gäller att identifiera sensitiva effektmått och hur sensitivitet på kort och lång sikt kan ökas (Lipsey & Cordray, 2000).

Eventuella ändringar i val av effektmått (6b)

För att undvika selektiv rapportering när det förekommer flera utfallsvariabler bör man på förhand specificera vilka utfall man avser

att mäta. Det kan emellertid vara nödvändigt att ändra utfallsmåtten, exempelvis om annan forskning tillkommer som visar att de valda utfallsmåtten inte är lämpliga eller att rekryteringsunderlaget är sämre än förväntat. Alla sådana ändringar måste rapporteras och motiveras. Detta för att visa att presentationen av resultat inte är selektiv, det vill säga att inte alla resultat presenteras.

Urvalsstorlek

Urvalsstorlek och statistisk styrka (7a)

En studie bör ha så många deltagare att det med hög grad av statistisk säkerhet (power) går att upptäcka en kliniskt meningsfull skillnad mellan experiment- och kontrollgrupp. Ju mindre skillnaden är mellan grupperna, desto större urval krävs för att upptäcka den. För att beräkna statistisk styrka krävs (1) den förväntade skillnaden mellan undersökningsgrupperna; (2) önskad alpha-koefficient (typ I-felnivå); (3) önskad statistisk styrka (beta eller typ II-felnivå) samt för kontinuerliga utfall: (4) standardavvikelse för mätningarna.

Det kan förekomma att skillnader i effektivitet mellan grupper är statistiskt säkerställda trots att ingen reell skillnad finns, alltså en **överskattning av skillnaden mellan grupper** (typ I-fel; falskt positiv). Det motsatta kan också förekomma, att skillnaden inte är statistiskt säkerställd trots att ett av de testade alternativen i verkligheten är effektivare, alltså en **underskattning av sambandet mellan positivt utfall och en behandlingsintervention** (typ II-fel, falskt negativ). Typ I-fel kan kontrolleras genom att ändra signifikansnivån från exempelvis .05 till .001. Att kontrollera typ II-fel är svårare eftersom det förutsätter att tillräckligt många personer deltar.

I rapporteringen bör det beskrivas hur beräkningen av statistisk styrka gått till; vilket primärt utfall beräkningen görs på, alla uppgifter som ingått i beräkningen samt urvalsstorlek i varje grupp. Detaljer om förväntade avhopp och bortfall vid uppföljningar bör också rapporteras. Det måste också rapporteras om den statistiska styrkan är tillräckligt hög. Den förväntade urvalsstorleken bör också rapporteras så att läsaren kan värdera om målsättningen uppnåd-

des. CONSORT avråder från post-hoc-analyser där man använder resultaten från undersökningen för att beräkna statistisk styrka i efterhand. Om urvalsstorleken avviker från det som planerats (t.ex. på grund av dålig tillströmning av deltagare) så ska det framkomma. Det förekommer att man i studier beräknar statistisk styrka efter hand för att avgöra huruvida studien ska stoppas eller om man ska förlänga inklusionsperioden. Generellt är urvalen i randomiserade studier små. I rapporteringen från studier med litet urval konkluderas ofta felaktigt att interventionsgrupperna inte var jämförbara, när det egentligen är för få deltagare för att kunna dra en sådan slutsats. Förmodligen är det relativt vanligt att det hade gått att upptäcka en skillnad om det hade varit ett större urval.

Avbruten inkludering och motivet till det (7b)

I många undersökningar pågår rekrytering av deltagare under en lång period. Författarna bör uppge hur många och vilka analyser som har gjorts under studiens gång samt om dessa var planerade från början. Om det efter hand visar sig att någon av interventionerna som studeras fungerar påtagligt sämre eller bättre än alternativet kan det av etiska skäl vara nödvändigt att avsluta undersökningen i förtid. En formell ”stoppregel”, det vill säga en gräns för när en studie bör avslutas i förtid, behöver därför utarbetas från början.

Randomisering

Steg i en typisk randomiseringsprocess är (1) en procedur för att generera en slumpmässig fördelningssekvens; (2) en metod för att dölja fördelningen (t.ex. numrerade förseglade kuvert) samt (3) implementering. Det rekommenderas att det inte är samma personer som tar fram fördelningssekvensen som också genomför randomiseringen i praktiken.

Metod för att generera slumpmässiga fördelningssekvenser (8a)

I rapporten ska det framgå hur deltagarna fördelats mellan grupperna. Därmed kan läsarna värdera sannolikheten för selektions-

bias. Det vanligaste är att använda en generator- eller slumptalstabell. Den slumpmässiga fördelningen innebär att det inte går att förutsäga vilken grupp deltagarna hamnar i och att sannolikheten är densamma för alla deltagare att hamna i den ena eller den andra gruppen. I somliga studier fördelas deltagarna avsiktligen ojämnt mellan grupperna eftersom man vill skaffa sig mer erfarenhet av den nya metoden eller begränsa kostnaderna för studien. Exempelvis kan proportionen experiment- och kontrollgrupp vara 2:1. Det ska dock noteras att en ökning i antalet personer i den ena gruppen bara marginellt påverkar statistisk power.

Restriktioner i randomiseringen (8b)

I stora urval om flera hundra personer är det som regel lätt att uppnå jämförbara grupper med hänsyn till kända och okända påverkansfaktorer. I mindre urval kan det vara lämpligt med restriktioner för att uppnå balans mellan grupperna när det gäller storlek på grupperna och karakteristika. Blockförsök innebär att försöksenheterna delas in i grupper (block) efter någon eller några bakgrundsvariabler. Därefter sker en slumpmässig fördelning till olika behandlingar inom varje block. Blockförsök säkrar att antalet individer i de olika grupperna blir i enlighet med det förutbestämda antalet, till exempel för varje block av åtta individer går fyra till experimentgruppen och fyra till kontroll. Vid användning av blockrandomisering behöver författaren rapportera hur blocken genererades samt antal individer som randomiseras till varje block.

Procedurer för att implementera den slumpmässiga fördelningen (9)

Det är viktigt att varken forskare, professionella eller klienter känner till vilken grupp en viss klient kommer att fördelas till innan baslinjemätningen är avslutad. Syftet är att förhindra selektionsbias. Det finns exempel på professionella som manipulerat randomiseringsprocessen eftersom de ”visste” vilken typ av behandling deras klient behövde.

Implementering av randomisering (10)

I en RCT ska deltagare rekryteras utifrån studiens inklusionskriterier. Nästa steg är att informera om studien och få samtycke till deltagande. En basmätning (förmätning) görs innan grupptilldelningen avslöjas. Det är önskvärt att randomiseringen genomförs centralt eller av en ”tredje part”. Om detta inte är möjligt kan numrerade kapslar eller förseglade kuvert användas. Det förutsätter att kuverten inte är genomskinliga och att proceduren är övervakad. Kuverten ska öppnas i rätt ordning efter att deltagarens namn och andra detaljer skrivits upp på kuvertet.

Maskering

Om aktuellt, hur doldes fördelning till grupperna (11a)

Det är önskvärt att personer som är inblandade i en effektutvärdering inte känner till vilken grupp deltagarna tillhör (s.k. blinding eller maskering) eftersom de kan påverkas av den kunskapen. Brist på maskering kan leda till under- eller överskattning av behandlingseffekterna. Till exempel kan deltagare reagera mer positivt på en behandling eftersom de vet att den är ny eller på grund av att datainsamlare (omedvetet) uppmuntrar undersökningspersoner vid prestationstest. I studier av psykosociala metoder är det dock ofta omöjligt att hålla de olika behandlingsalternativen dolda. Även om det inte löser problemet kan ett sätt vara att innan behandlingen administreras undersöka hur deltagarna uppfattar interventionen. Om deltagare i de olika grupperna har samma förväntningar antyder det att interventionerna i varje fall i förväg uppfattades som likvärdiga.

Om aktuellt, beskriv likheter mellan interventioner (11b)

Även om deltagare, professionella och datainsamlare inte informerats om vad som är experiment- respektive kontrollbetingelse kan det under vissa omständigheter vara enkelt att lista ut på grund av det sätt som det administreras, dess smak (om läkemedel) eller annat. Så långt som möjligt bör detta beskrivas.

Statistiska metoder

Statistiska metoder för att analysera primära och sekundära utfall (12a)

Rapporteringen behöver precisera vilka statistiska metoder som använts i varje analys. Beskrivningen ska vara så informativ att en kunnig person med tillgång till originaldata ska kunna göra om analyserna.

Gemensamt för nästan alla analysmetoder är att de ger ett estimat av behandlingseffekten som utgörs av kontrasten mellan utfallet i grupperna som jämförs. Konfidensintervall för behandlingseffekterna bör rapporteras eftersom dessa indikerar variationsbredden för osäkerhet när det gäller den sanna behandlingseffekten. Resultaten kan också redovisas med hjälp av statistisk signifikans, där *p*-värdet representerar sannolikheten för att en eventuell skillnad mellan grupperna uppstått av slumpen och inte beror på en sann skillnad mellan interventionerna. Faktiska *p*-värden (t.ex. $p = 0.003$) bör rapporteras och inte bara huruvida de över- eller understiger en viss nivå (t.ex. $p < 0.05$).

CONSORT betonar att analysmetoder ska användas där data förutsätts vara oberoende, det vill säga en observation per deltagare. Att använda multipla observationer från en och samma deltagare som oberoende data är ett allvarligt fel. I alla analyserna bör varje deltagare bara räknas en gång.

Kompletterande analyser (12b)

De metoder som använts i kompletterande analyser ska redovisas, till exempel subgruppsanalyser. Ett fel är att göra separata analyser av olika subgrupper, jämföra *p*-värden för respektive grupp och om den ena är signifikant och inte den andra dra slutsatsen att det finns en skillnad mellan grupperna.

Även i en randomiserad studie kan det finnas skillnader vid förmätningen i viktiga egenskaper hos deltagarna. Dessa initiala skillnader kan kontrolleras med hjälp av multipel regressionsanalys. Idealt anges motiv och typ av justerande analys redan i protokollet och bör inte baseras på att det finns statistiskt säkerställda skillnader

initialt (jfr punkt 15). Författare bör klargöra vilka variabler som justerades, hur kontinuerliga variabler hanterats, och specificera om analyserna var planerade eller motiverades av data.

Resultat

Undersökningsgrupp

Antal personer som randomiserats, fått behandling och analyserats (13a)

Ett flödesschema över hur undersökningsgruppen ser ut genom hela processen rekommenderas starkt: hur många som randomiserats till varje grupp, som fått den planerade behandlingen, som fullföljt behandlingen och som inkluderats i analyserna av resultat (se figur 1 för ett exempel). Denna information är avgörande för att kunna värdera om de som fullföljt utvärderingen är representativa för alla deltagare som randomiserats.

Det är viktigt att skilja mellan bortfall som beror på att personer inte gått att få tag på eller att personer har exkluderats av forskaren för att de inte motsvarade inklusionskriterierna och de personer som aktivt hoppat av behandlingen eller utvärderingen.

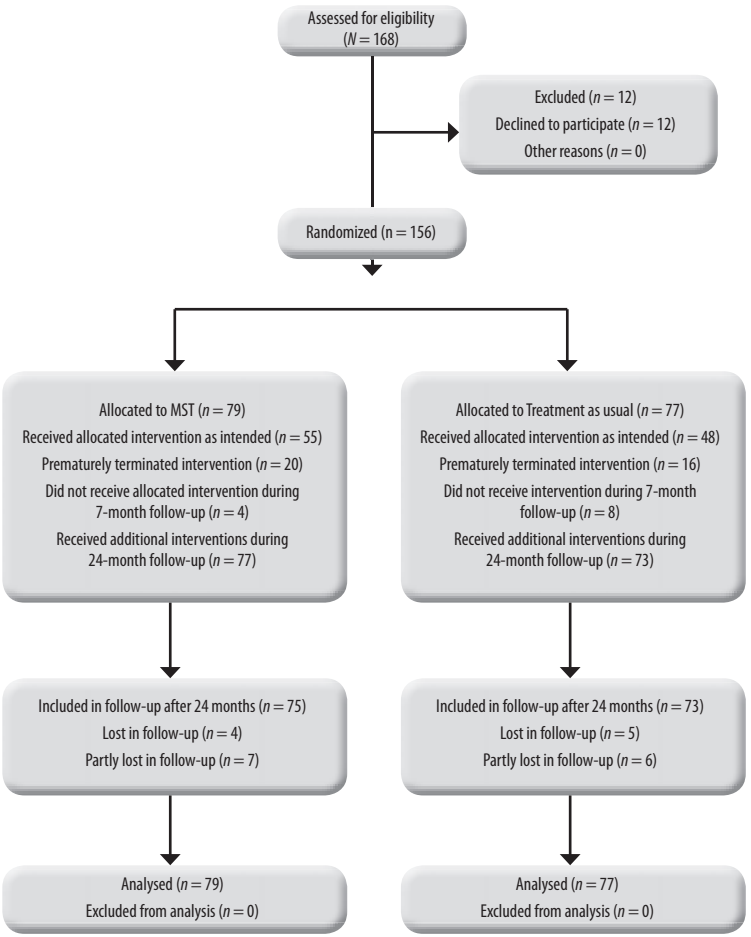
Bortfall efter randomisering och orsaker till det (13b)

Detta kan framgå av flödesschemat, men beskrivningar av eventuella avhopp, motiveringar till exkludering och andra avvikelser från planeringen bör ingå i rapporteringen om det är möjligt.

I figur 18:1 ges ett exempel på hur information om bortfall kan se ut i en artikel. Notera att figuren anger både ”lost in follow-up” och ”partly lost in follow-up”. Det senare syftar på att en uppgiftslämnare fallit bort men inte samtliga.

Rekrytering

Rekrytering av deltagare i en RCT tar sin utgångspunkt i en klart definierad målgrupp som interventionen är utformad för. Gruppen bör vara representativ för den population som man önskar gene-



Figur 18:1. Exempel på flödesschema över kontaktade, inkluderade, bortfall och analyserade (från Andrée Löfholm, Olsson, Sundell & Hansson, 2009).

ralisera resultaten till. Inklusions- och exklusionskriterier används för att på förhand definiera urvalet och därmed göra det möjligt att säga någonting om för vilka behandlingen har effekt och för vilka den inte har det. Målgruppen identifieras genom att använda reliabla och valida bedömningsinstrument eller intervjuer som fokuserar på målgruppens problematik. För att försäkra sig om ett homogent urval kan inklusionen baseras på en diagnos. Därmed reduceras san-

nolikheten för att multiproblem (komorbiditet) bidrar till att öka variationen i behandlingsutfall. En sådan avgränsning kan emellertid reducera resultatens användbarhet i ordinarie praktik där hjälpsökande ofta har multiproblem. Redan i rekryteringsfasen kan det ske en selektion beroende på om deltagarna självrekryteras (uppsöker vård frivilligt) eller inte.

Datum som definierar rekryteringsperiod och uppföljning (14a)

För att kunna placera en studie i ett historiskt sammanhang behövs information om var och när studien genomförts och under vilken tidsperiod deltagarna rekryterades. Det är viktigt bland annat eftersom behandlingsinterventioner kontinuerligt utvecklas, vilket kan påverka de reguljära interventioner som kontrollgruppen eventuellt får under undersökningsperioden. Att rapportera hur lång tid det tar att rekrytera deltagarna kan vara av intresse för andra forskare. Det bör på förhand rapporteras hur många och hur långa uppföljningar som planerades. Om uppföljningstiden skiljer sig åt mellan experiment- och kontrollgrupp ska detta rapporteras, gärna kompletterat med minimum, maximum och median. Det är också möjligt att lägga in uppföljningstid som ett kovariat i analyser av behandlingseffekt (t.ex. Amlund-Hagen, Ogden & Bjørnebekk, 2011).

Om aktuellt, varför utvärderingen avslutades i förtid (14b)

Det händer att utvärderingar avslutas i förtid, exempelvis eftersom skadliga effekter upptäckts i endera undersökningsgruppen eller eftersom experimentinterventionen visar på påtagliga effekter jämfört med kontrollgruppen. När en utvärdering avbryts i förtid eller förlängs ska de förhållanden som ligger bakom beslutet rapporteras. Det bör också eventuellt framgå vem som fattade beslutet.

Baslinjedata

Tabell över demografiska och kliniska kännetecken för varje grupp (15)

Randomiseringen är ingen garanti för att grupperna är ekvivalenta i alla avseenden vid förmätningen (baslinjemätning). Därför bör

grupperna presenteras med avseende på demografiska och relevanta kliniska kännetecken i en tabell, så att läsaren ges möjlighet att bedöma hur lika grupperna är. Detta är särskilt viktigt när inklusion baseras på data som avser primära utfall som problembelastning eller diagnos. Enligt CONSORT är det överflödigt och ologiskt att signifikanstesta skillnader mellan grupperna eftersom randomiseringen garanterar att eventuella skillnader beror på slumpen. I stället bör jämförelser vid förmätningen baseras på bedömningar av den prognostiska styrkan hos de variabler som mäts och hur mycket grupperna skiljer sig åt.

Antal analyserade (Intention-to-Treat eller Treatment of the Treated) (16)

En viktig fråga i randomiserade studier är hur många som inkluderas i analyserna och hur man ska förhålla sig till dem som avbryter behandlingen eller som aldrig börjar över huvud taget. Uttrycket ”intention-to-treat” (ITT) indikerar att samtliga personer som randomiserats till de olika behandlingsalternativen ingår i analyserna, oavsett om de hoppat av, bara gått vissa delar av behandlingen eller fullföljt den. Därmed bevaras randomiseringen. Den tidigare vanligaste analysmodellen ”treatment on the treated” (TOT) innebär att bara de som fullföljt behandlingen och som deltar i alla uppföljningar ingår i analysen. Nackdelen med att bara inkludera dem som fullföljt är att avhopp i både experiment- och kontrollgrupp kan vara selektivt, vilket leder till att randomiseringen blir komprometterad. Om exempelvis en av interventionerna lyckas bättre än den andra med att behålla deltagarna i behandling kan skillnaden i resultat mellan interventionerna bero på att den ena interventionen inte lyckats behålla deltagarna med den svåraste problematiken. I dag rekommenderar CONSORT ITT eftersom den analysmetoden bevarar randomiseringen och därmed indikerar bättre intern validitet. Samtidigt medför ITT ofta en underskattning av effekterna eftersom alla inte fått behandling.

Det händer också att individer på felaktiga grunder inkluderas

till studier, till exempel individer som inte uppfyller inklusionskriterierna, som aldrig deltagit i behandling eller som fått en annan behandling än den som planerades. Dessa deltagare ska trots det inkluderas i sina ursprungliga grupper enligt ITT, annars hotas randomiseringen. Det förekommer att dessa deltagare exkluderas ur analyser varvid modellen kallas ”modified intention to treat”. En sådan studie ska kallas icke-randomiserad.

Deltagare som avbryter medverkan eller som inte besvarar vissa datainsamlingsinstrument kan bara ingå i ITT-analyser om deras utfallsmått imputeras (med hjälp från annan insamlad information). CONSORT avråder från att använda enklare former av imputering som att använda det senast insamlade måttet för individer som saknar värden (”last measure carried forward”).

CONSORT 2010 har släppt kravet på ITT-analyser och i stället rekommenderas tydliga beskrivningar av exakt vilka och hur många individer som ingått i varje analys. Huruvida man bör välja ITT- eller TOT-analys avgörs av syftet med studien. Vid effektstudier med litet bortfall kan en TOT ge svar på frågan om behandlingen fungerar under optimala förhållanden. Men generellt sett lämpar det sig bättre med ITT-analyser, eftersom de genererar resultat som är tillämpbara på reguljär verksamhet där det är vanligare att klienter hoppar av behandlingen (Wright & Sim, 2003). I somliga studier kan det vara aktuellt att genomföra båda sortens analyser, till exempel i studier där många individer hoppar av behandlingen och det därmed blir stora skillnader mellan resultaten från TOT respektive ITT-analyserna.

Resultat och estimat

Resultat anges för primär- och sekundärmått och varje grupp (17a)

Resultaten för varje effektmått ska rapporteras i form av antal deltagare, medelvärden och standardavvikelser för respektive grupp samt kontrasten mellan grupperna, det vill säga effektstorleken.

För att indikera precision bör dessutom konfidensintervallet rapporteras för kontrasten mellan grupperna. I randomiserade studier

uttrycks huvudeffekter som kontrasten mellan den genomsnittliga skillnaden mellan experiment- och kontrollgrupp. Ett sådant punktestimat av behandlingseffekten ger inte någon heltäckande bild av individuella variationer när det gäller förändring. Somliga förändras mycket och andra mindre under en behandling. Om den statistiska styrkan är tillräckligt hög kan man uppnå en tillräckligt säker uppskattning av den genomsnittliga programeffekten, men beräkningen ger inte information om variationen runt medelvärdet. För att kompensera bristerna med ett punktestimat är det lämpligt att för varje genomsnittsresultat också redovisa konfidensintervall (vanligen 95 procent). Enligt CONSORT bör resultat aldrig rapporteras uteslutande med *p*-värden.

Selektiv rapportering är ett utbrett och allvarligt problem eftersom det ger en snedvriden bild av interventioners effekter. Därför bör samtliga planerade analyser rapporteras och inte bara analyser som är statistiskt signifikanta eller ”intressanta”.

För binära effektmått ska både absoluta och relativa effektstorlekar anges (17b)

När de primära utfallsmåtten är binära (t.ex. återfall i kriminalitet) ska både relativ risk (oddskvot, relativ risk) och absolut effekt (skillnad i risk) anges tillsammans med konfidensintervall.

Tilläggsanalyser – t.ex. subgrupps- och justerade analyser (18)

Testning av interaktion handlar om att analysera skillnad i behandlingseffekt i jämförbara grupper. Sådana analyser kan vara viktiga eftersom huvudeffekter kan dölja viktiga undergruppseffekter, till exempel mellan yngre och äldre deltagare (t.ex. Ogden & Amlund-Hagen, 2008). Men om man genomför många signifikanstester ökar risken för slumpmässiga fynd. Därför menar CONSORT att subgruppsanalyser bör undvikas. Detta gäller särskilt jämförelser mellan subgrupper som inte är planerade från början och som genomförs post hoc. Resultat från sådana analyser bekräftas sällan av andra studier och har därför låg trovärdighet. Om subgruppsana-

lyser genomförs ska motivet till det anges, om det var planerat från början samt vilka och hur många subgrupper som undersökts. Det ska inte genomföras på grund av statistiskt säkerställda skillnader vid förmätning.

Om man i analyserna justerar för bakgrundsskillnader vid basmätning så bör både justerade och icke-justerade resultat presenteras. Dessutom bör man presentera huruvida kovariansanalyser genomförts. Dessa justeringar bör vara planerade i forskningsprotokollet. Justeringar som beror på signifikanta skillnader vid basmätning kan leda till otillförlitliga resultat eller skevheter i den estimerade behandlingseffekten.

Förekomst av skadliga effekter (19)

Skadliga och oavsiktliga effekter som de studerade interventionerna förorsakat ska rapporteras. Detta är avgörande information eftersom det kan finnas interventioner som fungerar för de flesta men som har biverkningar för andra. Om inga skadliga effekter identifierats är det också viktig information att rapportera.

Diskussion

Begränsningar (20)

CONSORT menar att diskussionsdelen i vetenskaplig rapportering ofta är full av retorik som stödjer författarnas resultat och i mindre utsträckning utgör en balanserad beskrivning av det som talar för respektive emot undersökningen och dess resultat. CONSORT föreslår därför en mer strukturerad diskussionsdel som består av följande: (1) kort sammanfattning av resultaten; (2) bedömning och möjliga förklaringar till resultaten; (3) jämförelse med resultat från andra studier; (4) undersökningens begränsningar samt metoder som använts för att minska och kompensera för dessa; (5) kort avsnitt som summerar kliniska och för forskningen relevanta implikationer. Ovanstående punkter kan används som underrubriker i diskussionen.

Generaliserbarhet – extern validitet (21)

Generaliserbarhet handlar om studiens externa validitet eller användbarhet, och rör i vilken grad resultaten kan antas gälla för andra kontexter. Extern validitet är inte absolut men kan värderas i förhållande till kännetecken hos deltagarna, miljön som studien utförs i, behandlingsalternativet som prövas och resultaten. Därför bör information om inklusionskriterier, miljö, behandlingen, utfall samt tidsperiod för rekrytering och uppföljning alltid finnas med i rapporteringen. Hur många som uppfyllt inklusionskriterierna men som inte ville vara med i studien kan vara en indikation på klienternas preferenser och hur attraktiv en intervention förefaller vara. Det samma gäller för kliniska preferenser, det vill säga i vilken grad professionella i reguljära verksamheter är intresserade av att börja använda metoden.

Tolkning (22)

Tolkning av resultat ska balansera för- och nackdelar samt relatera resultaten till annan relevant evidens. Det bör göras en grundlig litteratursökning efter liknande studier så att relevanta jämförelser kan göras. Systematiska översikter får gärna vara en del av diskussionen.

Annan information

Registreringsnummer och namn på registret (23 + 24)

Allt fler vetenskapliga tidskrifter kräver att forskare i förväg ska publicera sina projektplaner. Det finns flera motiv till detta. Ett är att andra forskare runt om i världen ska veta vad som pågår för att undvika onödiga replikationer. Ett annat är att påskynda spridning av resultat; den som hittar en relevant studie i ett vetenskapligt register kan kontakta den ansvariga forskaren och få underhandsbesked långt innan de vetenskapliga resultaten finns publicerade. Ett tredje motiv är att det minskar risken för selektiv presentation av resultat. Det finns två typer av selektion. Den första är att studieresultat inte publiceras i vetenskapliga tidskrifter (publication bias),

något som främst gäller studier som inte kan uppvisa skillnader mellan interventioner. Det andra är att bara vissa resultat publiceras, företrädesvis de som resulterat i signifikanta skillnader, medan resultaten utan effekt inte redovisas (Littell, 2008). Båda typer av selektion riskerar att framställa en viss intervention som effektivare än den i själva verket är.

För det samhällsvetenskapliga området finns www.ClinicalTrials.gov som är öppen för forskare från hela världen att registrera sina projektplaner i. För hälso- och sjukvårdsområdet finns WHO:s forskningsregister över randomiserade studier, www.who.int/ictrp/en.

Finansiering och andra former av stöd (25)

För trovärdighetens skull är det viktigt att forskare öppet redovisar alla finansiärer till effektutvärderingen. Andra former av stöd bör också rapporteras.

Andra rapporteringskriterier

CONSORT har haft ett stort inflytande på kraven på rapportering av randomiserade kontrollerade studier i vetenskapliga tidskrifter. Men man har fått kritik för att de flesta kriterier handlar om den interna validiteten och att de därför bör kompletteras med kriterier som också berör den externa validiteten (Glasgow, Lichtenstein & Marcus, 2003; Glasgow, Vogt & Boles, 1999). Trudeau, Mostofsky, Stuhr och Davidson (2008) menar att psykologisk och beteendemedicinsk forskning är mer metodologiskt sårbar än exempelvis farmakologisk forskning och att det därför behövs flera kriterier. *The Evidence-based Behavioral Medicine Committee of the Society for Behavioral Medicine* efterlyser till exempel en mer detaljerad rapportering av utbildning och handledning av terapeuter, inklusive hur mycket utbildning som är nödvändig för att genomföra behandling på ett kompetent sätt för att förhindra bristande behandlingstrohet. En sådan rapportering bör inte bara omfatta huruvida klienten har fått behandlingen som planerat, utan också om han eller hon följt

terapeutens rekommendationer (Davidson, Goldstein, Kaplan m.fl., 2003). De har också föreslagit att man ska rapportera om terapeuten har en preferens för vissa behandlingsmetoder. Dessutom vill de inkludera mått på behandlingstrohet och om behandlingen hållit samma kvalitet under hela studien.

Glasgow och kollegor (1999, 2003) har också påpekat att i jämförelse med medicinska interventioner är psykosociala och pedagogiska interventioner ofta mer problematiska att definiera och standardisera. De riktar sig till en bred målgrupp som dessutom ofta inte själva söker vård. Både i modellutvärderingar och i verksamhetsutvärderingar är det viktigt att beskriva vilka interventioner som kan nå en brett definierad målgrupp, accepteras att användas av flera organisationer såsom skolor och sjukvårdsenheter, som går att implementera framgångsrikt av en brett sammansatt personalgrupp med varierande utbildning och kompetens och som ger replikerbara och varaktigt positiva resultat (och minimalt negativa resultat) till en rimlig kostnad. Detta ställer krav på statistiska metoder som kan hantera enskilda såväl som kontextuella faktorer (d.v.s. flernivåanalys). Glasgow med kollegor (1999; 2003) sammanfattar sin modell med akronymen RE-AIM som står för att nå ut (*Reach*), effektivitet (*Efficacy/effectiveness*), införande (*Adoption*), implementering (*Implementation*) och vidmakthållande (*Maintenance*). De betonar att randomiserade studier inte bara ska beskriva interventionens effekter på individen utan också beskriva hur stor andel av målgruppen som får ta del av interventionen och hur generaliserbar interventionen är utanför forskningssituationen.

R:et i RE-AIM syftar till i vilken utsträckning som interventioner når sin målgrupp. I denna analys kan man identifiera de hinder som begränsar deltagandet för vissa grupper, till exempel att de inte remitteras eller att det finns andra sociala och miljömässiga hinder för deltagande. Glasgow (2009) illustrerar problemen med att nå målgruppen med evidensbaserade interventioner med ”halveringsregeln”. Bland de verksamheter som skulle kunna använda en ny intervention deltar 40–60 procent. Av personalen i dessa verk-

samheter deltar 40–60 procent i implementeringen. Av dem som interventionen riktar sig till erbjuds 40–60 procent, av dem accepterar 40–60 procent erbjudandet och av dem fullföljer 40–60 procent behandlingen. Av dem gynnas 40–60 procent på kort sikt och av dem som gynnas på kort sikt kvarstår effekterna på lång sikt med 40–60 procent. Det innebär att av den totala målgruppen som interventionen vänder sig till är det mellan 0,3 och 5 procent som får bra resultat på kort sikt och 0,1–3 procent på lång sikt. Det är med andra ord viktigt att följa upp hur stor andel av de verksamheter för vilka interventionen är avsedd som faktiskt använder interventionen och hur många personer som får behandlingen. Även om det handlar om grova uppskattningar illustrerar Glasgow problemet med att verkligen hjälpa dem som interventioner avser att hjälpa.

A:et i RE-AIM, anpassning, handlar om hur många representanter av en viss verksamhet som kommer att använda interventionen och hur representativa de är för samtliga verksamheter. Hur många som använder interventionen påverkas av hur väl programmet är anpassat till politiska och kulturella förhållanden.

M:et i RE-AIM, vidmakthållande, syftar på om interventionen fortsätter att användas över tid, det vill säga i vilken utsträckning som den blir en bestående rutin i organisationen. Enligt Rogers (1995) är det endast ett fåtal interventioner som behålls över tid oavsett hur goda resultaten är under en inledande period.

Glasgow med flera (2003) påpekar att endast ett av CONSORT:s 22 kriterier behandlar generalisering och extern validitet. De menar i motsats till bland andra Kellam och Langevin (2003) att man i stället för att börja med modellutvärderingar (eng. *efficacy*) bör omedelbart testa lovande insatser i verksamhetsutvärderingar (eng. *effectiveness*) så att man kan avgöra hur många av målgruppen som får behandlingen respektive gynnas av den. Glasgow med flera (2003) fokuserar på hur väl interventionen accepteras och fungerar i professionellt arbete efter att interventionen befunnits vara effektiv. Det handlar om hur väl insatser når målgruppen, men också i vilken utsträckning de implementeras och vidmakthålls av organisationer.

Fördjupningslitteratur

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., m.fl. (2010). CONSORT 2010 Explanation and elaboration: updated guidelines for reporting parallel group randomised trials. Research Methods & Reporting. *BMJ, 340:* c869 (doi: 10.1136/bmj.c869).

Nezu, A. M. & Nezu, C. M. (red.). (2008). *Evidence-based outcome research. A practical guide to conducting randomized controlled trials for psychosocial interventions.* Oxford: Oxford University Press.

Schulz, K. F., Altman, D. G., Moher, D. for the CONSORT group (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trial. *BMC Medicine, 8,* 18.

Referenser

Amlund-Hagen, K., Ogden, T. & Bjørnebekk, G. (2011). Treatment Outcomes and Mediators of Parent Management Training: A One-Year Follow-up of Children with Conduct Problems. *Journal of Clinical Child and Adolescent Psychology, 40,* 165–178.

Andrée Löfholm, C., Olsson, T., Sundell, K. & Hansson, K. (2009). Multisystemic therapy with conduct-disordered young people: stability of treatment outcomes two years after intake. *Evidence and Policy, 5,* 373–397.

APA (2001). *Publication manual of the American Psychological Association, 5th edn.* Washington DC, APA.

APA (2006). Presidential task force on evidence-based practice. Evidence-based practice in psychology. *American Psychologist, 61,* 271–285.

Chambless, D. L. & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66,* 7–18.

Cunningham, A. (2002). *Lessons learned from a randomized study of Multisystemic Therapy in Canada.* PRAXIS: Research from the Centre for Children and Families in the Justice System. London/Canada, Centre for Children and Families in the Justice System (www.lfcc.on.ca).

Davidson, K. W., Goldstein, M., Kaplan, R. M., Kaufmann, P. G., Knatterud, G. L., Orleans, C. T., m.fl. (2003). Evidence-based behavioral medicine: What is it, and how do we achieve it? *Annals of Behavioral Medicine, 26,* 161–171.

Des Jarlais, D. C., Lyles, C., Crepaz, N. & TREND Group (2004). Improving the Reporting Quality of Nonrandomized Evaluations of Behavioral and Public Health Interventions: The TREND Statement. *American Journal of Public Health, 94,* 361–366.

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., Vandenbroucke, J. P. for the STROBE Initiative (2008). The Strengthening the Reporting of Observational studies in Epidemiology (STROBE) statement:

guidelines for reporting observational studies. *Journal of Clinical Epidemiology, 61,* 344–349.
Glasgow, R. E., (2009). Critical measurement issues in translational research. *Research on Social Work Practice, 19,* 560–568.
Glasgow, R. E., Vogt, T. M. & Boles, S. M. (1999). Evaluating the public health impact of health promotion interventions: the RE-AIM framework. *American Journal of Public Health, 89,* 1322–1327.
Glasgow, R. E., Lichtenstein, E., Marcus, A. C. (2003). Why don't we see more translation of health promotion research to practice? Rethinking the efficacy-to-effectiveness transition. *American Journal of Public Health, 93,* 1261–1267.
Kazdin, A. (2008). Evidence-based treatment and practice. New opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *American Psychologist, 63,* 146–159.
Kellam, S. G. & Langevin, D. J. (2003). A framework for understanding "evidence" in prevention research and programs. *Prevention Science, 4,* 3, 137–153.
Lipsey, M. W. & Cordray, D. S. (2000). Evaluation methods for social interventions. *Annual review of psychology, 51,* 345–375.
Littell, J. (2008). Evidence-based or biased? The quality of published reviews of evidence-based practices. *Children and Youth Services Review, 30,* 1299–1317.
Moher, D., Schulz, K. F. & Altman, D. G. (2001). The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials. *BMC Medical Research Methodology, 1:2.*
Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., m.fl. (2010). CONSORT 2010 Explanation and elaboration: updated guidelines for reporting parallel group randomised trials. Research Methods & Reporting. *BMJ, 340*: c869 (doi: 10.1136/bmj.c869).
Ogden, T. & Amlund-Hagen, K. (2008). Treatment effectiveness of Parent Management Training in Norway: A randomized controlled trial of children with conduct problems. *Journal of Consulting and Clinical Psychology, 76,* 607–621.
Rogers, E. (1995). *Diffusion of innovation. 4th edn.* New York: Free Press.
Schulz, K. F., Altman, D. G., Moher, D. for the CONSORT group (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trial. *BMC Medicine, 8,* 18.
Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.
Trudeau, K. J., Mostofsky, E., Stuhr, J. K. & Davidson, K. W. (2008). Explanation of the CONSORT statement with application to psychosocial interventions. I: A. M. Nezu & C. M. Nezu (red.). *Evidence-based outcome research. A practical guide to conducting randomized controlled trials for psychosocial interventions* (s. 25–44). Oxford: Oxford University press.
Wright, C. C., Sim, J. (2003). Intention-to-treat approach to data from randomized controlled trials: A sensitivity analysis. *Journal of Clinical Epidemiology, 56*, 833–842.

Kapitel 19 | Knut Sundell & Terje Ogden

Effektutvärderingar och framtiden

Bokens första kapitel inleddes med ett citat om att de flesta experiment sker i vardaglig verksamhet och utan forskarmedverkan. I dag ökar antalet effektutvärderingar i Sverige, men från en mycket låg nivå. Vår förhoppning är att den här boken ska bidra till att nya effektutvärderingar håller hög vetenskaplig kvalitet. Dåliga effektutvärderingar medför i bästa fall ett slöseri med resurser och de medverkandes tid och engagemang. I sämsta fall kan det leda till att ineffektiva eller skadliga interventioner fortsätter att användas eller sprids. Att få en ny intervention att bli använd är en utmaning i sig. Att utmönstra en ineffektiv intervention kan vara en ännu större utmaning.

Effektutvärderingar handlar om tillämpad forskning; de genomförs med ambitionen att resultaten ska komma till användning. Eftersom det i huvudsak handlar om interventioner som finansieras med skattemedel, fattas besluten om vilka interventioner som ska prioriteras av politiker eller kommunalt anställda tjänstemän. I den beslutsprocessen konkurrerar forskningen med andra informationskällor som exempelvis ideologi, ekonomi och kultur. Den forskning som håller högsta kvalitet har bäst möjlighet att komma till användning.

Tekniken för att genomföra effektutvärderingar med maximal

trovärdighet utvecklas ständigt. Det i sin tur förändrar innebörden av vad som är hög kvalitet; det som idag är metodologiskt avancerat kommer förmodligen att vara standard om några år. Och det som ansågs vara god kvalitet tidigare är inte alltid det i dag. Tidigare accepterades exempelvis artiklar för publikation av klusterrandomiserade studier, som inte hanterat beroende mellan dem som ingick i samma kluster. Det går knappast i dag. Andra kvalitetskriterier som allt mer kommit att betonas är beräkning av statistisk styrka, en utförlig beskrivning av rekryteringen av undersökningsgruppen (inklusive orsaker till att personer avböjt respektive avbrutit medverkan), imputering av bortfall samt dokumentation och analys av behandlingstrohet, mediatorer och moderatorer. I dag räcker det inte att visa vilken intervention som är mest effektiv. Forskaren behöver också visa hur stark effekten är, om den förändras över tid, om interventionen använts som tänkt, om den fungerar lika för olika målgrupper samt om den tänkta förändringsfaktorn är relaterad till utfallet.

I detta avslutande kapitel behandlas först två frågor som inte behandlats i de tidigare kapitlen: dels vad som ska utvärderas, dels vägar för att öka antalet effektutvärderingar. Sist lyfter vi fram fem kriterier som vi uppfattar som särskilt viktiga att eftersträva i en effektutvärdering.

Vad ska utvärderas?

Även om det vore önskvärt är det inte rimligt att utvärdera alla psykosociala och pedagogiska interventioner, bland annat av ekonomiska skäl. Inom medicinen, som prioriterat effektutvärderingar sedan lång tid, uppges exempelvis endast 20 procent av behandlingsmetoderna vara evidensbaserade (Nordenström, 2009). Av de 103 interventioner som svensk socialtjänst uppger att de använder i öppenvård för barns psykiska hälsa (Socialstyrelsen, 2009) är tio utvärderade. Om det behövs minst två av varandra oberoende utvärderingar av hög kvalitet för att dra säkrare slutsatser om evidens,

krävs ytterligare cirka 200 effektutvärderingar. Om varje utvärdering kostar mellan tre och fem miljoner, blir den totala kostnaden 600 till 1 000 miljoner!

Alla interventioner behöver inte utvärderas experimentellt. I en provokativ artikel gjorde Smith och Pell (2003) en systematisk översikt av om fallskärmar är effektiva för att förhindra ”allvarliga trauman i relation till utmaning från gravitation”. Eftersom de inte hittade någon RCT föreslog de att de som kräver att allt måste utvärderas experimentellt ska delta i en dubbelblind, randomiserad, placebokontrollerad crossover-studie av fallskärmsanvändning. Man kan på goda grunder anta att ingen vill delta i en sådan studie! Exemplet med fallskärmar åskådliggör att allt inte ska utvärderas i kontrollerade experiment. Samtidigt är det viktigt att notera att på de flesta pedagogiska, samhällsvetenskapliga och medicinska områden visar sig inte effekter av interventioner direkt, utan det normala är att effekter är svaga och först uppenbaras efter en längre tid. När det gäller att hoppa från flygplan med eller utan fallskärm visar sig effekterna omedelbart och det finns få modererande faktorer att ta hänsyn till. När det gäller psykosociala behandlingar är regeln att en rad modererande faktorer kan misstänkas, till exempel ålder, motivation och förekomsten av ett socialt nätverk, för att nämna några.

Eftersom effektutvärderingar är både tidskrävande och dyrbara, är det endast ett begränsat antal interventioner som kan utvärderas på detta sätt. Därmed behövs någon form av prioritering vad gäller vilka interventioner som ska utvärderas.

Prioriterade interventioner att utvärdera

Vad som är prioriterade forskningsbehov i ett visst samhälle varierar med förekomst av problem, utbyggd vård och omsorg, politiska prioriteringar och vetenskapliga prioriteringar, för att nämna några faktorer. Så vitt känt finns ingen enhetlig prioritering. Resonemang om när screening är etiskt motiverat kan fungera som en inspiration. Enligt Juth och Munthe (2012) kan screening motiveras när de problem som ska åtgärdas är vanligt förekommande samt

har allvarliga konsekvenser för individen. Allvarliga konsekvenser kan vara hög grad av skada (på hälsan), låg ålder när problematiken startar (eftersom det kan innebära livslång problematik) samt sannolikheten att problematiken kommer att förvärras.

Det skulle innebära att interventioner som bör prioriteras att utvärdera är de som behandlar vanligt förekommande problem och som har allvarliga konsekvenser för individen. Andra interventioner som bör prioriteras är de som vänder sig till barn och unga med en problematik som riskerar att förvärras. Vidare bör interventioner prioriteras som individen har begränsad möjlighet att välja bort (låg autonomi), exempelvis olika former av tvångsvård.

Därutöver finns även andra kriterier för prioritering. Dit hör interventioner som har evidens i systematiska översikter från andra samhällssystem och kulturer. Interventioner som framstår som lovande i pre-post-studier utan jämförelsegrupp och som genererat stora effektstorlekar kan också prioriteras.

För samtliga förslag ovan gäller dessutom att interventionerna är väldefinierade, gärna beskrivna i manualer, så att interventionen kan komma att användas av andra på rätt sätt.

Det finns också fall då en utvärdering knappast är motiverad. Ett exempel är att importera och utvärdera interventioner som redan utvärderats i andra länder och som saknar effekt eller som till och med är skadliga.

Slutligen bör nya interventioner ha genomgått flera inledande test innan de blir föremål för en effektutvärdering med upprepad mätning och kontrollgrupp, eftersom de senare är både dyrbara och tidskrävande. Interventioner bör först ha testats och befunnits acceptabla för både klienter och professionella och det behövs även vetenskapligt stöd för att klienter som får interventionen förbättras.

Prioriterade interventioner att utvärdera

- De som berör många och riktas mot problem som har allvarliga konsekvenser för individen.
- De som berör barn och unga.
- De som individer har små möjligheter att välja bort.
- De som har evidens i systematiska översikter från andra samhällssystem och kulturer.
- De som framstår som lovande i exempelvis pre-post-studier utan jämförelsegrupp och som genererat stora effektstorlekar.
- De som är väldefinierade (t.ex. har manual).

Sätt att främja effektutvärderingar

I dag finns ett tydligt ökat nordiskt intresse för effektutvärderingar. En utmaning på kort sikt är att öka antalet forskare med kompetens att bedriva effektutvärderingar.

På sikt är det viktigt att fler doktorander tränas i effektutvärderingar inom ramen för sina doktorandarbeten. Förutsättningen för att utbilda doktorander är dock att det finns handledare som behärskar metodiken. En nyckeluppgift är således att fortbilda empiriskt verksamma forskare. Ett annat sätt att möta bristen på handledarkompetens är tvärvetenskapliga institutioner där forskare från discipliner med större erfarenhet av effektutvärderingar ingår (t.ex. omvårdnadsforskning, psykologi och medicin). Ett tredje sätt är ett internationellt utbyte som ger svenska forskarstuderande möjlighet att utbildas i länder som har den efterfrågade kompetensen.

Ett annat problem är att forskningsråden inte har prioriterat effektutvärderingar inom de psykosociala och utbildningsvetenskapliga områdena. Bland annat har forskningsråden varit negativa till att finansiera implementering av de interventioner som ska utvärderas. Om kommuner måste finansiera försöksverksamheter, vars värde är oklart, riskerar de att slösa bort skattemedel. För kommuner blir det då rationellt att låta andra kommuner finansiera nödvändiga försöksverksamheter för att i efterhand och utan kostnad utnyttja den vunna kunskapen. Detta ”free rider-problem” behöver

lösas antingen genom att forskningsråden också bidrar med medel till implementering eller att regering eller myndigheter har särskilda medel för detta.

Nordiskt samarbete

Ett sätt att hantera problemet med kostsamma effektutvärderingar är att genomföra dem som nordiska samarbetsprojekt, eftersom de nordiska samhällssystemen liknar varandra kulturellt, ekonomiskt och samhällsmässigt. Det gäller framför allt interventioner som vänder sig till kvantitativt sett små målgrupper som det annars kan ta lång tid att inkludera tillräckligt många individer av. Om ett utvecklat nordiskt samarbete också visar att effekterna är snarlika i de olika länderna, öppnar det för att en ny intervention inte nödvändigtvis måste utvärderas i alla nordiska länder, utan att det kan räcka med ett land.

Ett exempel på ett sådant samarbete är en effektutvärdering av multifunktionell behandling för ungdomar på institution och i närmiljö (MultifunC). MultifunC bygger på empirisk forskning om det som kännetecknar effektiva institutioner samt om metoder som har visat positiva effekter för ungdomar med allvarliga beteendeproblem. MultifunC har implementerats på tre enheter inom Statens institutionsstyrelse samt på något fler i Norge. Det betyder att antalet ungdomar som kommer att genomgå behandlingen under en överskådlig tid är begränsat. Genom att använda samma design, datainsamlingsinstrument med mera ges en möjlighet att slå samman data från Norge och Sverige och på det sättet få ökad statistisk styrka för att upptäcka om det finns skillnader jämfört med traditionell behandling.

Några kvalitetskriterier för effektutvärderingar

En effektutvärdering blir inte bättre än sina sämsta delar. En dåligt planerad datainsamling kan exempelvis öka risken att undersökningspersoner hoppar av från utvärderingen. Det i sin tur minskar

den statistiska styrkan, vilket kan äventyra utvärderingens möjlighet att säkert besvara frågeställningar. En studie som planeras ha för låg statistisk styrka är oetisk, eftersom det är resursslöseri och utsätter undersökningspersonerna för onödig ansträngning.[1]

Varje kapitel i denna bok ger viktiga bidrag för planering och genomförande av en kvalitativt avancerad effektutvärdering. Det betyder inte att boken är heltäckande. Det finns områden som den inte behandlar. Dit hör exempelvis det sista viktiga steget i en effektutvärdering – hur forskaren kan göra för att den erhållna kunskapen ska kunna spridas och komma till användning.[2]

Avslutningsvis behandlar vi fem kriterier som vi uppfattar som särskilt viktiga att försöka uppfylla i en effektutvärdering.

Randomiserad kontrollerad utvärdering

Det bästa sättet att säkra att resultaten från en effektutvärdering får inflytande är att använda en randomiserad design. En RCT har en viktig fördel mot andra utvärderingsdesigner. Om randomiseringen fungerar skapas två grupper som i teorin är likvärdiga med hänsyn till bakgrundsegenskaper och som endast skiljer sig åt i att den ena gruppen får interventionen och den andra inte. Om interventionsgruppen vid eftermätning eller uppföljningsmätning i genomsnitt har förbättrats mer än jämförelsegruppen, ökar sannolikheten att det beror på interventionen, det vill säga att den interna validiteten är hög. Det betyder inte att experiment är det enda sättet att fastställa orsakssamband, men det är den bästa metoden för att säkerställa intern validitet; att det finns ett kausalt samband mellan intervention och utfall. Utan intern validitet blir exempelvis den externa validiteten mindre intressant. Därför bör en experimentell design vara det första alternativet att överväga vid effektutvärderingar.

---

1 Däremot kan resultat från utvärderingar som misslyckats med att uppnå tillräckligt stor undersökningsgrupp vara värdefulla i metaanalyser eftersom de då ingår i den sammanlagda undersökningspopulationen från flera primärstudier.

2 T.ex. Fixsen, Naoom, Blasé, Friedman & Wallace, 2005; Roselius & Sundell, 2008.

Det hävdas ibland att RCT är omöjligt att använda i samhällsvetenskaplig forskning. Den invändningen kommer ofta från forskare som har en bakgrund i andra forskningstraditioner och som saknar egna erfarenheter av RCT. I dag finns uppskattningsvis 10 000 RCT av sociala interventioner internationellt och cirka hundra svenska (Socialstyrelsen, 2011).

Randomiserade effektutvärderingar har kritiserats för att inte svara på frågan om *hur* interventioner fungerar. Det är sant att de inte ger ett säkert svar på den frågan. Men det är möjligt att med hjälp av mediator- och moderatoranalys testa om de antagna mekanismerna är statistiskt relaterade till avläsbara effekter (MacKinnon, 2008). I den norska utvärderingen av den familjebaserade interventionen Parent Management Training visade analyser att barns beteende medierades av kvaliteten i föräldraskapet (Ogden & Amlund-Hagen, 2008; Amlund-Hagen, Ogden & Bjørnebekk, 2011). Mediator- och moderatoranalyser ger inte underlag för starka slutsatser om huruvida en viss mekanism påverkar kausalt, men de kan visa att en mekanism *inte* påverkar om den inte är statistiskt relaterad till utfallet. Det första steget i en RCT är sålunda att testa om det finns genomsnittliga skillnader mellan de grupper som undersökts (om interventionen fungerar) för att därefter undersöka om interventionen fungerar bättre för någon undergrupp (moderatoranalys) och om de antagna mekanismerna för förändring är relaterade till förändringen (mediatoranalys).

Samtidigt ställer experimentella studier speciella krav (t.ex. Sanson-Fisher, Bonevski, Green & D'Este, 2007). Det kan exempelvis vara svårt att få professionella att acceptera att deras klienter kan komma att lottas till en väntelista. En annan svårighet i verksamhetsbaserade studier är att säkra att interventionen genomförs som den ska, med tillräcklig behandlingstrohet. En tredje svårighet handlar om kontaminering; ju längre uppföljningstid som används, desto större är risken att kontrollgruppen får ta del av den experimentella interventionen. Det gäller framför allt om kontrollgruppen utgörs av en väntelista.

Ibland är det omöjligt att använda RCT. I det fallet är ett icke-randomiserat experiment med jämförelsegrupp det bästa alternativet. Denna typ av effektutvärdering har också sina förespråkare och kan under optimala förhållanden generera ett tillförlitligt resultat (t.ex. Shadish, Clark & Steiner, 2008).

Sammantaget betyder det att även om randomiserade studier har begränsningar, har alternativen i allmänhet fler. Vi menar därför att forskare som avser att utvärdera en intervention bör utgå från att det går att genomföra en RCT. Först när det är klarlagt att den designen inte kan användas bör andra typer av designer övervägas (jfr Weisburd, 2010).

Standardiserad design

Syftet med effektutvärderingar är att få kunskap om vilka effekter en viss intervention har så att det kan vägleda professionellt socialt, psykologiskt och pedagogiskt arbete. Det är först när det finns flera effektutvärderingar som ger samstämmiga resultat som en säkrare kunskap finns om interventionens effekter. Det säkraste sättet att få den kunskapen är med hjälp av systematiska översikter. Den systematiska översikten följer samma vetenskapliga metodik som primärforskning. Den ska redovisa frågeställning, metod, genomförande och tolkning på ett sådant sätt att andra forskare kan reproducera resultaten. Det ställer krav om att de primärstudier som inkluderas tydligt redovisar relevanta fakta om hur effektutvärderingen genomförts och vilka resultat som erhållits. För att skapa konsensus om hur det ska ske har bland annat CONSORT utvecklats.

Vikten av att effektutvärderare börjar använda en mer standardiserad metodik åskådliggörs i en artikel av Andrée Löfgren, Brännström, Olsson och Hansson (under tryckning) där de granskar effektutvärderingar av multisystemisk terapi. Av 13 RCT fanns inte två som hade samma jämförelsegrupp, uppföljningstid och primära effektmått. Det skapar problem för den systematiska översikten.

Adekvat kontrollgrupp

Avgörande för tolkningen av en effektutvärdering är valet av kontrollgrupp; utan kontrollgrupp är det svårt eller omöjligt att avgöra hur stark effekt interventionen haft. Men valet av kontrollgrupp påverkar i allra högsta grad resultaten.

Grovt sett finns det tre typer av kontrollgrupper: (1) ingen intervention (inkl. väntelista); (2) en annan aktiv intervention och (3) en blandning av aktiva interventioner. Inom medicinsk forskning används också placebointervention, men det är mindre vanligt inom det psykosociala området. Valet av intervention påverkar teoretiskt och praktiskt storleken på de effekter som kan förväntas. En obehandlad kontrollgrupp resulterar oftast i en starkare effekt för experimentgruppen än om kontrollgruppen fått en aktiv behandling. En nackdel med en obehandlad kontrollgrupp är att bristen på intervention kan förmå dessa personer att vid sidan om utvärderingen söka hjälp på annat håll, vilket kan fördunkla slutsatserna om interventionens effekter.

Inom medicinsk forskning används ofta en standardbehandling som jämförelsegrupp. Det är den intervention som det finns konsensus om att vara den mest effektiva. Men om det saknas en standardbehandling, som ofta är fallet inom psykosocialt och pedagogiskt arbete, uppstår problem. Framför allt om det finns flera interventioner som parallellt används som standardbehandling och som klumpas ihop i en effektutvärdering. I det fallet blir det svårt att översätta resultaten till praktisk verksamhet (Burns, 2009). Om en viss arbetsplats framför allt använder standardbehandling A men inte alls standardbehandling B och C, går det inte att avgöra om det är standardbehandling A, B eller C eller samtliga eller en kombination av dem som varit sämre än den intervention som testats specifikt. Ett exempel på hur det kan se ut är den svenska utvärderingen av multisystemisk terapi (MST), där de 77 ungdomarna i kontrollgruppen fick sju olika interventioner och där några inte fick någon intervention alls.

I det fall det saknas standardbehandling kan en rekommenda-

tion vara att använda en obehandlad kontrollgrupp (t.ex. väntelista), eftersom den effekt som då erhålls är lättare att tolka. Ett annat alternativ till kontrollgrupp är att välja en standardbehandling som tidigare utvärderats mot en obehandlad kontrollgrupp, något som ger möjlighet att extrapolera effekten. Ett tredje alternativ är att välja den standardbehandling som har starkast teoretiskt stöd.

Det finns tillfällen när det är svårt att skapa en obehandlad kontrollgrupp. Det gäller särskilt när rekryteringen av kontrollgruppen sker via en organisation där personerna redan är aktuella; dels kan personerna aktivt ha sökt sig till en viss intervention och inte vara motiverade för alternativ, dels kan det finns en juridisk skyldighet att erbjuda behandling. Det finns dock exempel där en ”obehandlad” kontrollgrupp rekryterats via annonser i regional- och rikstäckande dagspress. Scheffer Lindgren och Tengström (2011) annonserade efter kvinnor som utsatts för våld av sin manliga partner men som inte hade någon pågående kontakt med socialtjänsten eller kvinnojour på grund av våldet. Över 200 kvinnor kom på detta sätt att ingå i en icke-randomiserad kontrollerad utvärdering, utöver de cirka 350 som tillhörde en aktiv behandlingsgrupp. Visserligen fanns skillnader mellan grupperna vid förmätningen, men med hjälp av regressionstekniker kan initiala skillnader i teorin kontrolleras – givet att de relevanta selektionsfaktorerna har dokumenterats (jfr Shadish m.fl., 2008).

Transparens

Många vetenskapliga tidskrifter kräver att forskare redovisar egna finansiella intressen i den intervention eller motsvarande som en artikel behandlar, till exempel om forskaren bedriver kommersiell utbildningsverksamhet av den intervention som utvärderas. Eisner och Humphreys (2011) visar att effektstorlekarna var högre i effektutvärderingar av brottspreventiva interventioner när forskarna hade kommersiella intressen i den intervention de utvärderat. Det betyder inte nödvändigtvis att vissa forskare fuskar, utan det kan handla om kognitiv bias (”tunnelseende”) eller högre programtro-

het eftersom programkonstruktören medverkar (Petrosino & Soydan, 2005). Inte heller betyder det att en forskare som utvärderar en egen intervention automatiskt blir mindre trovärdig. Vetenskaplig trovärdighet handlar i första hand om transparens – att så exakt som möjligt redovisa hur utvärderingen genomförts.

> ***Exempel på bristande transparens***
> Eisner (2009)
>
> I en metaanalys av ett namngivet preventionsprogram drog författarna slutsatsen att interventionen var effektiv. Det resultatet, oavsett hur korrekt det är, kan ifrågasättas, eftersom en av den publicerande tidskriftens två redaktörer också ingick i preventionsprogrammets rådgivande kommitté och ansvarade för en utvärdering av preventionsprogrammet i USA. Vidare implementerade en av författarna till metaanalysen programmet i Tyskland, ingick i preventionsprogrammets rådgivande kommitté och hade en anknytning till det kommersiella företag som distribuerar programmet i Tyskland. Inget av detta redovisades av vare sig författare eller redaktörer.

Ett sätt att öka transparens och motverka verkliga eller inbillade hot om snedvridning är att forskare i förväg registrerar sina projektplaner (se även kapitel 18). För det samhällsvetenskapliga området finns www.ClinicalTrials.gov där forskare från hela världen kan registrera sina projektplaner.

Registerdata för uppföljning

En anledning till att effektutvärderingar är kostsamma är att det kan behövas relativt lång tid för att samla en tillräckligt stor undersökningsgrupp, som dessutom behöver följas en längre tid för att effekterna ska kunna avläsas. Det medför att tre år i allmänhet är ett minimum från planering till rapport. Vid efter- och uppföljningsmätningar behöver de personer som ingår i studien i allmänhet besökas personligen för datainsamlingen.

Ett sätt att minska resursbehoven är att använda registerdata som effektmått. Det är inte alltid möjligt, men i vissa fall finns relevanta

registerdata tillgängliga. Ett exempel är lagföringsregister över kriminalitet när en utvärdering handlar om att förebygga kriminalitet. Andra register som kan vara aktuella är register över hälso- och sjukvård och försörjningsförmåga.

Användning av registerdata kan eliminera behovet av att följa upp deltagare personligen vid efter- och uppföljningsmätning. I stället kan resurserna koncentreras på att träffa undersökningspersonerna vid förmätningen och samla in information om relevanta bakgrundsvariabler, samt att registrera programtrohet för interventionerna under interventionsfasen. Att undersökningspersoner som byter bostadsort kan vara kostnadsdrivande åskådliggör en studie av social barnavård där 31 procent av 239 personer bytte bostadsort inom en treårsperiod (Sundell, 2002).

Utöver att registerdata kan spara resurser uppstår inte något eller endast ett begränsat bortfall (individer kan emigrera). En annan fördel är att det kan vara mer etiskt vid långtidsuppföljningar, eftersom personer som deltagit i interventionen kan tycka att det är obehagligt att påminnas om ett tidigare problem. En nackdel är att registerdata ofta har en viss eftersläpning. Registerdata har också brister, liksom alla typer av data. Även om bortfallet i allmänhet är lågt, är det ofta störst för sociala marginalgrupper, exempelvis skolbetyg bland barn i dygnsvård. Trovärdigheten kan dessutom ibland ifrågasättas, exempelvis avser brottsanmälningar endast en del av all brottslighet.

Registerdata används framgångsrikt inom medicinsk och samhällsvetenskaplig forskning i form av retrospektiva fall-kontroll- och kohortstudier (t.ex. Vinnerljung, Franzén & Danielsson, 2007; Vinnerljung, Hjern & Lindblad, 2006). Däremot har registerdata sällan använts som utfallsmått i prospektiva experimentella utvärderingar, något som det är hög tid att ändra på. Ett känt exempel är *Cambridge-Somerville Youth Study,* som startade 1935 i Boston, USA. Syftet var att förebygga kriminalitet bland pojkar i slummen. Pojkarna lottades antingen till en grupp som fick stöd av socialarbetare under uppväxten eller till en grupp som inte fick det. När pojkarna följdes upp 30 år se-

nare uppgav två tredjedelar i behandlingsgruppen att projektet varit av avgörande betydelse för deras uppväxt (McCord, 2003). Registerdata gav dock en helt annan bild. Personerna i behandlingsgruppen hade oftare dömts för allvarlig kriminalitet, fler var alkoholiserade och fler hade allvarliga psykiska sjukdomar eller hade avlidit.

Referenser

Amlund-Hagen, K., Ogden, T. & Bjørnebekk, G. (2011). Treatment Outcomes and Mediators of Parent Management Training: A One-Year Follow-up of Children with Conduct Problems. *Journal of Child and Adolescent Psychology, 40,* 165–178.

Andrée Löfholm, C., Brännström, L., Olsson, M. & Hansson, K. (under tryckning). Treatment-as-usual in effectiveness studies: what is it and does it matter? *International Journal of Social Welfare.*

Burns, T. (2009). End of the road for treatment-as-usual studies? *The British journal of Psychiatry, 1958,* 5–6.

Eisner, M. (2009). Reply to the comments by David Olds and Lawrence Sherman. *Journal of experimental criminology, 5,* 215–218.

Eisner, M. & Humphreys, D. (2011). Measuring conflict of interest in prevention and intervention research: A feasibility study. I T. Bliesener, A. Beelmann & M. Stemmler (Red.). *Antisocial Behavior and Crime: Contributions of Developmental and Evaluation Research to Prevention and Intervention* (s. 165–180). Cambridge, MA: Hogrefe Publishing.

Fixsen, D. L., Naoom, S. F., Blasé, K. A., Friedman, R. M. & Wallace, F. (2005). *Implementation Research: A Synthesis of the Literature.* University of South Florida. http://www.fpg.unc.edu/~nirn/resources/publications/Monograph/pdf/Monograph_full.pdf

Juth, N. & Munthe, C. (2012). *The ethics of screening in health care and medicine. Serving society or serving the patient?* International library of ethics, law, and the new medicine 51. DOI 10.1007/978-94-007-2045-9_1. Springer Science+Business Media B.V. 2012

Littell, J. H. (2008). Evidence-based or biased? The quality of published reviews of evidence-based practices. *Children and Youth Services Review, 30,* 1299–1317.

McCord, J. (2003). Cures that harm: unanticipated outcomes of crime prevention programs. *Annals of the American academy, 587,* 16–30.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis.* Mahwah, NJ: Erlbaum.

Nordenström, J. (2009). *Evidensbaserad medicin i Sherlock Holmes fotspår.* Stockholm: Karolinska Institutet University Press.

Ogden,T. & Amlund-Hagen,K. (2008). Treatment effectiveness of Parent

Management Training in Norway: A randomized controlled trial of children with conduct problems. *Journal of Consulting and Clinical Psychology, 76,* 607–621.
Petrosino, A. & Soydan, H. (2005). The impact of program developers as evaluators on criminal recidivism: Results from meta-analyses of experimental and quasi-experimental research. *Journal of Experimental Criminology, 1*, 435–450.
Roselius, M. & Sundell, K. (Red) (2008). *Att förändra socialt arbete*. Stockholm: IMS och Gothia Förlag.
Sanson-Fisher, R. W., Bonevski, B., Green, L. W. & D'Este, C. (2007). Limitations of the randomized controlled trial in evaluating population-based health interventions. *Am J Prev Med*, *33*, 155–61.
Scheffer Lindgren, M. & Tengström, A. (2011). *Utvärdering av Socialtjänstens och Ideella kvinnojourers Insatser för Våldsutsatta kvinnor*. Stockholm: Karolinska Institutet.
Shadish, W. R., Clark, M. H., Steiner, P. M. (2008). Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments. *Journal of the American Statistical Association. 103,* 1334–44.
Smith, G. C. S. & Pell, J. P. (2003). Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. *BMJ, 327*, 1459–1461.
Socialstyrelsen (2009). *Socialtjänstens öppna verksamheter för barn och unga – en nationell inventering av metoder*. Stockholm: Socialstyrelsen.
Socialstyrelsen (2011). *Svensk och internationell forskning om sociala interventioners effekter*. Stockholm: Socialstyrelsen.
Sundell, K. (2002). *Familjerådslag i Sverige. Socialtjänstens fortsatta insatser till barn och föräldrar* (FoU-rapport 2002:3). Stockholms socialtjänstförvaltning: FoU-enheten.
Vinnerljung, B., Franzén, E. & Danielsson, M. (2007). Teenage parenthood among child welfare clients – a Swedish national cohort study. *Journal of Adolescence*, 30, 97–116.
Vinnerljung, B., Hjern, A. & Lindblad, F. (2006). Suicide attempts and severe psychiatric morbidity among former child welfare clients – a national cohort study. *Journal of Child Psychology and Psychiatry,* 47, 723–733.
Weisburd, D. (2010). Justifying the use of non-experimental methods and disqualifying the use of randomized controlled trials: challenging folklore in evaluation research in crime and justice. *Journal of Experimental Criminology*, 209–227.

Maria Bodin

Ordlista

Begrepp	*Förklaring*
Alternativ hypotes (H_1) (eng. *alternative hypothesis*)	Mothypotes, forskningshypotes. Vid ▶statistisk hypotesprövning är den alternativa hypotesen den som accepteras om ▶nollhypotesen visar sig vara falsk, d.v.s. då gruppernas medelvärden eller frekvenser skiljer sig så mycket att skillnaden sannolikt inte förklaras av slumpen.
Alternativkostnad (eng. *opportunity cost*)	Vid ekonomiska analyser; värdet av det alternativ som väljs bort till förmån för ett annat. Om man exempelvis arbetar 10 timmar i veckan frivilligt, vad skulle värdet vara om man i stället hade fått betalt för de timmarna?
ANOVA	▶Variansanalys.
Bedömningsbias (eng. *detection bias*)	Otillförlitlighet i resultat från en ▶effektutvärdering som uppstår när det finns systematiska skillnader för hur man bedömer utfall i studiegrupperna. Kan exempelvis uppstå när bedömare som stått för mätning av resultaten känner till vilken intervention deltagarna fått inom ramen för studien. Bedömaren kan skatta större förbättringar vid uppföljningsmätning för deltagare som fått den ▶intervention som antas vara den bästa. Aktuellt för utfallsmått som innefattar subjektiva bedömningar, medan risken är mindre för objektiva mått. Risken för bedömningsbias minskar med ▶blindning.

Begreppsvaliditet (eng. *construct validity*)
Den grad med vilken en skala eller ett test lyckas fånga det teoretiska begrepp som man avser mäta och inget annat. Empiriskt stöd för detta finns när testet ▶korrelerar med andra test som mäter liknande faktorer (konvergent validitet) och samtidigt inte har något samband med faktorer som är begreppsmässigt orelaterade (divergent validitet).

Behandlingsbias (eng. *performance bias*)
Otillförlitlighet i resultat från en ▶effektutvärdering som uppstår när det finns systematiska skillnader mellan studiegrupperna gällande olika behandlingsfaktorer. Det kan till exempel handla om att kontrollgruppen i högre grad än experimentgruppen söker behandling utanför studien.

Behandlingsgrupp (eng. *treatment group*)
▶Experimentgrupp.

Behandlingsmetod
▶Intervention.

Behandlingstrohet (eng. *treatment integrity, treatment fidelity*)
Den grad med vilken en ▶intervention genomförs på det sätt som föreskrivs i en teoretisk modell eller en manual. Behandlingstrohet kan omfatta flera dimensioner, t.ex. ▶följsamhet och kompetens.

Beroendevariabel (eng. *dependent variable*)
En ▶variabel som antas kunna påverkas av värdet på en ▶oberoende variabel. Inom ramen för en ▶effektutvärdering är beroendevariabeln den ▶utfallsvariabel som man vill påverka med den studerade ▶interventionen. ▶Oberoende variabel.

Bevisvärde (eng. *study quality*)
Den skattade vetenskapliga kvaliteten hos en enskild studie och dess förmåga att besvara frågan om en interventions effekter på ett utfallsmått på ett tillförlitligt sätt. En studie med hög risk för ▶bias har ett lägre bevisvärde. Bevisvärdet hos enskilda studier ligger till grund för bedömningen av det vetenskapliga underlagets sammantagna ▶evidensstyrka. Såväl bevisvärde som evidensstyrka för en intervention kan variera mellan utfallsmått.

Bias
Systematiskt fel som medför att resultat från en ▶effektutvärdering blir otillförlitligt. I ▶systematiska översikter bedöms graden av ▶risk för bias hos de ingående primärstudierna. ▶Selektionsbias, ▶Bedömningsbias, ▶Behandlingsbias, ▶Bortfallsbias, ▶Rapporteringsbias ▶Publikationsbias ▶Minnesbias.

Blindning
(eng. *blinding*)

Att inom ▶effektutvärderingar hålla deltagare, behandlare och dem som genomför mätningarna ovetande (blinda) inför deltagarnas grupptillhörighet (d.v.s. vilka deltagare som är i ▶experimentgrupp respektive i ▶kontrollgrupp). Syftet är att undvika ▶reaktivitet och ▶bedömningsbias vilket kan ge otillförlitliga resultat. Enligt ▶CONSORT-uttalandet från 2010 bör termerna ”singelblindad” och ”dubbelblindad” undvikas vid rapportering eftersom det inte är tydligt vad som avses. I stället bör det beskrivas vilka som hållits blinda. Blindning av såväl deltagare som behandlare och bedömare genomförs regelmässigt i läkemedelsstudier. I studier av psykologiska och sociala interventioner är det svårt att ”blinda” deltagarna och dem som ger interventionen.

Bonferronikorrektion

Korrigering av ▶signifikansnivån när man statistiskt testar flera ▶hypoteser. Signifikansnivån $p < .05$ gäller för ett enda test och risken för att hitta en statistiskt signifikant, slumpmässig skillnad ökar med antalet genomförda test. Bonferronikorrektion innebär att man justerar den godtagbara signifikansnivån genom att dela den ursprungliga nivån med antalet testade hypoteser. Vid fem genomförda test blir signifikansnivån efter korrektion $p < .01$ (0.05/5). Korrektionen innebär att risken för ▶typ-I-fel minskar.

Bootstrapmetoden
(eng. *bootstrapping*)

Bootstrapping innebär att man gör upprepade slumpmässiga urval beträffande en variabel från ett existerande dataset (resampling). Kan användas för att skaffa ett ▶estimat av en ▶variabel (t.ex. dess ▶medelvärde eller varians) när data inte uppfyller antaganden om ▶normalfördelning och lika varians.

Bortfall
(eng. *attrition*)

Bortfall innebär att data från en mätning i en ▶longitudinell studie saknas helt eller delvis på grund av tekniskt missöde, att personen inte längre samtycker till deltagande i mätningen eller inte längre går att komma i kontakt med. Om bortfallet i en ▶effektutvärdering ser olika ut i de studerade grupperna uppstår risk för ▶bortfallsbias.

Bortfallsbias
(eng. *attrition bias*)
Otillförlitlighet i resultat från en ▶effektutvärdering som uppkommer då bortfallets storlek eller orsak skiljer sig mellan de studerade grupperna, exempelvis om deltagare med initialt större problemtyngd fallit bort i högre utsträckning i en av grupperna. Till skillnad från många andra hot mot den ▶interna validiteten minskar risken för bortfallsbias inte med ▶randomisering.

Box-plot
”Låddiagram”. Grafisk framställning av en variabel där observationerna rangordnats och där ▶medianvärde och ▶percentiler framgår. En box-plot kan användas för att identifiera ▶extremvärden.

Campbell Collaboration
Ett internationellt nätverk av bland annat forskare som genomför, publicerar och regelbundet uppdaterar ▶systematiska översikter avseende socialt arbete, kriminalvård och utbildning, för att få fram resultat som kan bidra till en ▶evidensbaserad praktik.

Cochrane Collaboration
Ett internationellt nätverk av bland annat forskare som genomför, publicerar och regelbundet uppdaterar ▶systematiska översikter avseende hälso- och sjukvårdsinterventioner, för att få fram resultat som kan bidra till en ▶evidensbaserad praktik.

Cohens d
En ▶effektstorlek som räknas ut genom att man subtraherar en grupps ▶medelvärde från en annan grupps medelvärde och dividerar med gruppernas genomsnittliga ▶standardavvikelse. Vanligtvis motsvarar d = 0.20 en liten effekt, d = 0.50 en medelstor effekt och d = 0.80 en stor effekt av interventionen.

Confounder
▶Ovidkommande variabel.

CONSORT
CONSORT Group (Consolidated Standards of Reporting Trials) producerar och ger ut riktlinjer för hur ▶randomiserade kontrollerade studier bör rapporteras i vetenskapliga tidskrifter. Flera tidskrifter kräver i dag att forskare som sänder in manuskript visar i vilken utsträckning de följt riktlinjerna i sin rapportering genom att bifoga CONSORT-gruppens ”checklista”.

DfBeta
DfBeta används för identifikation av multivariata ▶extremvärden. Beskriver graden av förändring i en regressionskoefficient (betakoefficient) när en person inkluderas respektive inte inkluderas i en multipel regressionsanalys. Grafiskt motsvarar detta graden av förändring i regressionslinjens lutning.

Dikotom variabel
(eng. *dichotomous variable, binary variable*)
En variabel med två kategorier som är varandra uteslutande (exempelvis kön). En dummyvariabel är en dikotom variabel med värdena 0 och 1 som används i regressionsmodeller för att indikera frånvaro eller närvaro av någon egenskap eller händelse.

Direkta kostnader
(eng. *direct costs*)
Vid ekonomiska analyser; de resurser som används vid planering, implementering, leverans och fortsatt användande av en intervention.

Dos-respons-relation
(eng. *dose-response relationship*)
Ett samband mellan en given behandlingsdos och dess effekt på ett ▶utfallsmått. Vanligen avses ett positivt samband, d.v.s ju mer behandling, desto större effekt.

Drop-out
Deltagare som i förtid avslutar medverkan i en studie. Om termen ska användas i rapportering bör man specificera om den avser avhopp från behandlingen/interventionen, ▶bortfall vid uppföljningsmätningarna eller båda dessa.

Effectiveness trial
▶Verksamhetsutvärdering.

Effekt
(eng. *effect*)
Den observerade relationen mellan en ▶intervention och ett ▶utfall. Beskrivs som en ▶effektstorlek.

Effektivitet
(eng. *effectiveness*)
Att använda tillgängliga resurser på bästa sätt.

Effektstorlek
(eng. *effect size*)
Statistisk term som beskriver storleken och riktningen på effekten av en viss intervention. Exempel på effektstorlekar är ▶Cohens d, ▶oddskvot, ▶riskkvot och ▶number needed to treat.

Effektutvärdering
(eng. *outcome study, impact evaluation, intervention study*)
Studie vars syfte är att (1) värdera effekten av en intervention, (2) fokusera på de berörda personernas förbättrade hälsa och psykosociala situation, (3) beskriva slutresultatet av interventionen samt (4) vars resultat är avsedda att användas som underlag för beslut som fattas av dem som deltar i vård, omsorg och pedagogisk verksamhet.

Effektvariabel
(eng. *outcome variable*)
▶Utfallsvariabel.

Efficacy trial
▶Modellutvärdering.

Efterföljande ekonomisk analys
(eng. *retrospective economic analysis*)
En ekonomisk analys gjord på en studie som från början designats för att besvara enbart kliniska frågor.

Efterföljande resurser
(eng. *downstream resources*)
Resurser som används vid en senare tidpunkt.

Ekonomiska kostnader
(eng. *economic costs*)
▶Totala kostnader, inklusive pengar spenderade på produktion och ▶alternativkostnader.

Empirisk prövning
(eng. *empirical test*)
Vetenskapligt insamlande av data i syfte att testa en ▶hypotes.

Estimat
(eng. *parameter estimate*)
Skattning av en ▶parameter utifrån ett ▶urval.

Evidens
(eng. *empirical evidence*)
Något som bedöms tyda på att ett visst förhållande gäller. Inom ramen för begreppet ▶evidensbaserad praktik utgör evidens det vetenskapliga stödet för en insats eller åtgärds effekt på ett ▶utfallsmått. Vetenskapligt stöd är inte något absolut utan handlar om grad av tilltro till vilket vetenskapligt stöd som finns för åtgärdens eller insatsens effekt på ett givet utfallsmått. I vetenskapliga sammanhang utgör evidens empiriskt insamlade belägg för eller emot ett antagande. Normerna för hur dessa belägg ska samlas in och därmed för vad som utgör evidens för en intervention kan variera såväl mellan som inom kunskapsområden.

Evidensbaserad praktik
(eng. *evidence-based practice*)
En medveten och systematisk användning av bästa tillgängliga ▶evidens, tillsammans med den professionellas expertis samt den berörda personens situation, erfarenhet och önskemål, för beslut om interventioner till enskilda personer. Hur informationen från de olika kunskapskällorna vägs samman bestäms av det nationella och lokala sammanhanget, till exempel genom lagstiftning, riktlinjer och tillgängliga resurser.

Evidensstyrka
(eng. *quality of evidence*)
Evidensstyrkan beskriver tillförlitligheten hos det sammanlagda vetenskapliga underlaget för en viss slutsats. Det saknas enhetliga system för att bedöma evidensstyrka. Flera svenska myndigheter använder idag det internationellt utarbetade evidensgraderingssystemet ▶GRADE i ▶systematiska översikter om diagnostik och interventioners effekter.

Exklusionskriterier
(eng. *exclusion criteria*)
Förhållanden eller egenskaper som forskaren i förväg har bestämt ska exkludera personer från medverkan i en studie (t.ex. annan samtidig behandling).

Experimentell studie
(eng. *experimental study*)
En ▶studiedesign där forskaren medvetet introducerar en ▶intervention för att avläsa dess effekter. Kontrasteras ofta mot ▶observationsstudier. Innefattar vanligtvis åtminstone två mätningar – före och efter interventionens start. Om experimentet är ▶kontrollerat innefattar designen även en jämförelsegrupp som får en annan intervention än experimentgruppen. Om den kontrollerade studien dessutom fördelar deltagare slumpvis till grupper, är det en ▶randomiserad, kontrollerad studie.

Experimentgrupp
(eng. *experimental group*)
Den grupp deltagare i en utvärdering som administreras den intervention som testas och vars utfall jämförs med en ▶kontrollgrupp eller en ▶jämförelsegrupp.

Extern validitet
(eng. *external validity*)
Avser i vilken utsträckning en studies resultat går att generalisera till andra personer i målgruppen, platser och sammanhang än dem som var aktuella inom studien.

Experimentvariabel
▶Oberoende variabel.

Extremvärde
(eng. *outlier*)
Ett värde som avviker från flertalet övriga värden i en datamängd. En univariat outlier avviker med värdet för en ▶variabel, en multivariat outlier avviker med värdet för flera variabler.

F-kvot
(eng. *F-value*)

F-kvoten (efter Ronald A. Fisher) är ett testvärde som oftast används inom ▶variansanalys för att avgöra om det finns en statistiskt signifikant skillnad mellan de undersökta gruppernas ▶medelvärden. F-kvoten visar om variationen mellan grupperna är tillräckligt mycket större än variationen inom grupperna för att förkasta ▶nollhypotesen. Man säkerställer statistisk signifikans med hjälp av en F-tabell som visar hur stor F-kvoten måste vara, givet de aktuella ▶frihetsgraderna, för att man ska förkasta nollhypotesen. Motsvarande testvärden med liknande funktion finns för andra typer av statistiska test, exempelvis t-värdet och χ^2-värdet.

Fall-kontroll-studie
(eng. *case-control study*)

En typ av ▶observationsstudie som vanligtvis innebär att forskaren jämför förekomsten av en förmodad orsaksfaktor (t.ex. rökning) hos personer som har det studerade problemet (fall = t.ex. personer med lungcancer) med förekomsten av samma faktor hos friska personer (kontroll = t.ex. personer utan lungcancer). För varje fall utses en eller flera kontroller.

Fallstudie
(eng. *case study*)

En studie som omfattar en enda individ.

Fasta kostnader
(eng. *fixed costs*)

Kostnader som måste betalas även om en organisation inte producerar någonting.

Finansiella kostnader
(eng. *financial costs*)

Bokföringskostnaden för en vara eller tjänst, vilket normalt är det belopp som ursprungligen betalades och inte ▶alternativkostnaden.

Flernivåanalys
(eng. *multilevel analysis*)

Kallas även hierarkiska modeller eller analyser med mixade effekter eller slumpeffekter. Används när det finns en hierarkisk struktur eller klusterstruktur i data, vilket innebär att observationerna inte är oberoende av varandra. Data från exempelvis skolor tenderar att vara klustrade eftersom elever inom en skola liknar varandra mer än de liknar resten av populationen skolelever.

Forskningshypotes

▶Alternativ hypotes.

Forskningsplan
(eng. *study protocol*)

En noggrant genomtänkt skriftlig forskningsplan är till hjälp för genomförandet av en ▶effektutvärdering och för att man i slutändan ska kunna besvara forskningsfrågan. Den efterfrågas också vid forskningsetisk prövning. Introduktionen i en forskningsplan beskriver den planerade utvärderingens teoretiska och empiriska bakgrund, samt studiens syften, frågeställningar och hypoteser. Metoddelen beskriver planerad rekrytering av deltagare, beräkning av ▶statistisk styrka och deltagarantal, mätmetoder och planerade statistiska analyser. Planen innefattar även en tidsplan, uppskattade kostnader, bilagor samt en referenslista.

Fraktiler
(eng. *fractiles*)

Fraktiler är värden som delar upp ett rangordnat datamaterial i delar med lika många observationer i varje, exempelvis ▶medianvärdet som delar det i två lika stora delar, kvartiler (fyra delar), deciler (tio delar) och percentiler (hundra lika stora delar).

Frekvens
(eng. *frequency*)

Hur ofta något inträffar eller hur många individer som tillhör en viss kategori.

Frihetsgrader
(eng. *degrees of freedom*)

Antalet frihetsgrader används vid ▶statistisk hypotesprövning för att avgöra den kritiska gränsen för när ett testvärde är tillräckligt stort för att visa på en signifikant skillnad mellan grupperna. Sätten att räkna ut frihetsgrader skiljer sig mellan olika test men är ofta ett uttryck för antalet individer, antalet variabler och antalet grupper som ingår i analysen.

Följsamhet
(eng. *adherence compliance*)

En behandlares efterlevnad av en manual eller en behandlingsmodell. Avser även en studiedeltagares, klients eller patients efterlevnad av en tilldelad ▶intervention eller behandling.

Försöksledareffekt
(eng. *experimenter effect*)

Avser att försöksledarens förväntningar eller egenskaper påverkar den egna studiens resultat.

Genomsnitt
(eng. *arithmetic mean*)

▶Medelvärde.

Genomsnittliga kostnader
(eng. *average costs*)

▶Totala kostnader dividerat med summan av produktionen, till exempel antalet behandlingstimmar som produceras under året.

GRADE

GRADE är ett internationellt utarbetat evidensgraderingssystem för bedömning av tillförlitligheten hos vetenskapliga underlag som ligger till grund för slutsatser om diagnostik och interventioners effekter. Bedömningen av tillförlitlighet görs per ▶utfallsmått och indelas i fyra nivåer. Utgångspunkten är att om underlaget består av ▶randomiserade kontrollerade studier anses det finnas ett starkt vetenskapligt underlag. Poängavdrag kan sedan göras utifrån fem faktorer som bedöms minska tillförlitligheten: bristande studiekvalitet (risk för olika typer av ▶bias för utfallsmåttet i de bedömda primärstudierna), bristande överförbarhet (förutsättningarna eller utfallsmåtten i primärstudierna brister i relevans för exempelvis svenska förhållanden), bristande samstämmighet (olika studiers resultat för ett utfallsmått pekar i olika riktning), bristande statistisk precision (den sammanvägda effekten är osäker på grund av ett litet antal deltagare eller inträffade händelser i studierna) samt risk för ▶publikationsbias.

Grå litteratur
(eng. *grey literature*)

Vetenskapliga studier som inte har publicerats i etablerade vetenskapliga medier, exempelvis myndighetsrapporter, vissa avhandlingar och examensarbeten.

Hawthorne-effekt
(eng. *Hawthorne effect*)

Reaktion hos deltagare i en studie som ändrar sitt beteende på grund av att de blir uppmärksammade snarare än på grund av den studerade ▶interventionen.

Hypotes (vetenskaplig)
(eng. *hypothesis*)

Ett antagande om hur till exempel en specifik ▶intervention påverkar ett visst utfall, som testas i en ▶statistisk hypotesprövning. ▶Nollhypotes, ▶Alternativ hypotes.

Hävstångsvärden
(eng. *leverage values*)

Värden som kan användas för att identifiera multivariata ▶outliers vilka kan ändra en regressionsmodells ”fit” till data.

Immateriella kostnader
(eng. *intangible costs*)

Kostnader för exempelvis obehag, smärta, oro och besvär.

Implicit kostnad
(eng. *implicit cost*)
Den kostnad som kommer av att man valt ett alternativ framför ett annat. Till exempel om man reser världen runt under ett års tid, så är den implicita kostnaden lönen som man inte får under året.

Imputering
(eng. *imputation*)
Att i en statistisk undersökning ersätta värden som saknas med uppskattade värden.

Indirekta kostnader
(eng. *indirect costs*)
Värdet av vårdgivarnas egna resurser som förbrukats i syfte att förse klienterna med interventionen.

Informerat samtycke
(eng. *informed consent*)
En persons samtycke till att delta i en vetenskaplig studie efter att han eller hon fått muntlig och skriftlig information om studiens syfte och genomförande. Informerat samtycke från deltagare är ett krav för att studien ska godkännas i forskningsetisk prövning.

Inklusionskriterier
(eng. *inclusion criteria*)
De betingelser som krävs för att en person ska kunna ingå i en undersökning. Exempelvis i en studie av behandling för alkoholberoende är ett inklusionskriterium diagnosticerat alkoholberoende. Att ▶informerat samtycke getts är ett obligatoriskt inklusionskriterium. Deltagande i undersökningen kräver dessutom att inget av ▶exklusionskriterierna föreligger. Liknande urvalskriterier används även för en ▶systematisk översikt och avser då vilka villkor som en publicerad undersökning ska uppfylla för att den ska inkluderas i översikten.

Inkrementell analys
(eng. *incremental analysis*)
En analys av differensen mellan olika åtgärder avseende kostnader eller effekter.

Innehållsvaliditet
Inom testteori huruvida ett mätinstrument fångar alla relevanta aspekter av det man avser att mäta. Ett exempel är om ett begåvningstest som enbart fångar logisk förmåga har en sämre innehållsvaliditet än ett som fångar även verbal och spatial förmåga. Innehållsvaliditet kan även avse överensstämmelse mellan det en forskare önskar att studera och de mätinstrument som ingår i effektutvärderingen. Expertbedömningar är det vanligaste sättet att avgöra om innehållsvaliditeten är tillgodosedd.

Inre konsistens
(eng. *internal consistency*)
Även kallad intern konsistens. Graden av samstämmighet mellan item i en skala eller ett test som är tänkt att fånga samma teoretiska begrepp. ▶Reliabilitet.

Intern konsistens ▶Inre konsistens.

Insats ▶Intervention.

Intention to treat-analys (ITT) ITT-analys innebär att man vid analys av data från en effektstudie (a) inkluderar samtliga deltagare som fördelats till grupperna, (b) låter deltagarna behålla sin grupptillhörighet oavsett om de avbrutit interventionen eller fått annan intervention än den avsedda samt (c) gör uppföljningsmätning med samtliga deltagare. Eftersom bortfall är vanligt i longitudinella studier kräver det sista kriteriet oftast någon form av ▶imputering, d.v.s. uppskattning av de värden som saknas. ITT-analys anses vara att föredra framför ▶TOT-analys för resultatbearbetning i ▶effektutvärderingar.

Interaktion När ett statistiskt samband mellan en ▶oberoende variabel och en ▶beroendevariabel påverkas i riktning eller styrka av en tredje variabel. Till exempel om en behandling fungerar olika bra för män och kvinnor finns en interaktion mellan behandlingseffekt och kön. Kön är då en ▶moderatorvariabel för behandlingseffekten.

Interbedömarreliabilitet (eng. *inter-rater reliability, inter-rater agreement*) Graden av samstämmighet mellan två eller flera personers bedömning vid användning av en och samma mätprocedur, till exempel ett instrument för strukturerad beteendeobservation. ▶Reliabilitet.

Intern validitet (eng. *internal validity*) Avser i vilken utsträckning ▶interventionen snarare än systematiska fel (▶bias) kan förklara en effekt i en effektutvärdering. Ett viktigt hot mot den interna validiteten är att undersökningsgrupperna inte är lika vid utvärderingens start, till exempel att en grupp är mer motiverad till behandling (▶selektionsbias). ▶Randomisering kontrollerar för flera hot mot intern validitet.

Intervallskala (eng. *interval scale*) Typ av ▶mätskala. Variabler på intervallskalenivå är sådana där skillnader mellan skalsteg (till skillnad från ▶ordinalskala) har samma innebörd oavsett var på skalan man befinner sig. Till skillnad från variabler på ▶kvotskala har de på intervallskala inte en absolut nollpunkt. Ett exempel på en intervallskala är Celsiusskalan.

Intervention
(eng. *intervention*)
En medveten åtgärd för att åstadkomma en förändring och som syftar till att uppnå ett visst mål (t.ex. förebygga psykiska eller sociala problem) för en definierad ▶population. Interventioner sammanfattas i form av en skriftlig eller muntlig överförbar kunskap, och görs tillgängliga genom utbildning, undervisning, handledning eller självstudier.

Interventionsgrupp
(eng. *intervention group*)
▶Experimentgrupp.

Interventionsstudie
▶Effektutvärdering.

Intressekonflikt
(eng. *conflict of interest*)
Jävsituation. Avser här att en författare till en vetenskaplig rapport har finansiella eller andra personliga intressen i det som studerats, vilket kan ha påverkat resultaten.

Jämförelsegrupp
(eng. *comparison group*)
▶Kontrollgrupp.

Kapitalkostnader
(eng. *capital costs*)
Värdet av de kapitalresurser som har längre livslängd än ett år.

Kausalsamband
(eng. *causal relationship*)
Orsak–verkansamband. Om en variabel ska kunna antas orsaka förändring i en annan variabel krävs bland annat att en statistisk ▶korrelation föreligger, att den förmodade orsaken föregår utfallet i tid, samt att alternativa förklaringar till förändringen kan uteslutas.

Klusterrandomisering
(eng. *cluster-randomization*)
Typ av ▶randomisering där det inte är individer utan grupper av individer som fördelas slumpmässigt till interventions- och jämförelsegrupp, exempelvis skolor eller socialkontor. Designen kräver speciell metodik när man bestämmer storlek på undersökningsgruppen och under analysfasen. ▶CONSORT har speciella riktlinjer för rapportering av klusterrandomiserade studier som går utöver dem som gäller för randomiserade studier generellt.

Kohortstudie
(eng. *cohort study*)

En typ av ▶observationsstudie där en grupp personer med vissa definierade egenskaper (kohort) följs över tid. Vanligen studeras två eller flera undergrupper i kohorten som jämförs med avseende på förloppet över tid, till exempel återfall i kriminalitet eller sjukdom. En kohortstudie är vanligtvis ▶prospektiv men kan även vara ▶retrospektiv.

Konfidensintervall (KI)
(eng. *confidence interval*)

En statistisk term som anger säkerheten hos ett mätvärde. Ju smalare konfidensintervall, desto större är säkerheten. Exempelvis innebär ett 95-procentigt KI att om man skulle dra flera stickprov från en och samma ▶population och beräkna ett konfidensintervall för vart och ett av dessa, så skulle 95 procent av intervallen innefatta populationens sanna värde.

Kontamination
(eng. *diffusion*)

I en ▶kontrollerad studie den oavsiktliga administrationen (läckage) av den studerade interventionen till deltagare i kontrollgruppen.

Kontrafaktisk inferens
(eng. *counterfactual inference*)

I ▶kontrollerade studier något som är motsatsen till fakta (t.ex. vad som skulle ha hänt en viss grupp individer om de inte hade fått en speciell intervention). I allmänhet är förutsättningen för kontrafaktisk inferens att det finns en ▶kontrollgrupp som interventionsgruppen kan jämföras mot.

Kontrollerad studie
(eng. *controlled trial*)

En ▶experimentell studie där interventionsgrupp och ▶jämförelsegrupp följs över tid. En kontrollerad studie kan vara ▶randomiserad eller icke-randomiserad.

Kontrollgrupp
(eng. *control group*)

Den grupp i en kontrollerad studie som inte får den studerade interventionen och med vilken experimentgruppen jämförs. Kontrollgruppen kan vara obehandlad, ställas på ▶väntelista, få en overksam behandling (▶placebokontroll) eller den intervention som brukar ges i ordinarie verksamhet (▶traditionell behandling). Tidigare användes termen kontrollgrupp för att beteckna ett kontrafaktiskt alternativ som inte fått någon behandling och jämförelsegrupp som den som fått en annan behandling, till exempel standardbehandling.

Korrelation
(eng. *correlation*)
Statistiskt begrepp som anger att två variabler samvarierar numeriskt. Korrelationen kan vara positiv (ju mer av den ena, desto mer av den andra) eller negativ (ju mer av den ena, desto mindre av den andra). En korrelation innebär inte per automatik ett ▶kausalsamband, d.v.s. att den ena variabeln orsakar variationen i den andra.

Kostnads-intäktsanalys
(eng. *cost-benefit analysis*)
En teknik där både kostnaderna och effekterna uttrycks i monetära termer. Nettonuvärde används ofta för att sammanfatta resultatet.

Kostnads- och effektanalys
(eng. *cost-consequence analysis*)
Påminner om ▶kostnads- och effektivitetsanalys. I den redovisas kostnader i monetära termer och effekterna i effektenheter separat. Inget försök görs att summera resultatet.

Kostnads- och effektivitetsanalys
(eng. *cost-effectiveness analysis*)
En teknik där kostnaderna uttrycks i monetära termer och effekten i effektenheter. För denna metod krävs att endast en effekt studeras eller att effekterna kan vägas samman. Mått som kostnadseffektkvot eller kostnad per effektenhet används ofta för att sammanfatta resultatet.

Kriterievaliditet
(eng. *criterion validity*)
Ett mått på hur väl ett test eller mätinstrument samvarierar med ett visst uppsatt kriterium. Kriterievaliditet kan vara samtidig eller prediktiv, beroende på om testet och mätningen av kriterievariabeln görs vid samma tidpunkt eller vid olika tillfällen. Om det finns en hög ▶korrelation mellan testresultatet för ett riskbedömningsinstrument för återfall i brott och anmälning i brottsregister vid senare tidpunkt har instrumentet hög prediktiv validitet.

Kvalitativ studie
(eng. *qualitative study*)
En studie som syftar till förståelse av människors upplevelser och som inte besvarar frågor genom siffermässiga resultat utan genom tolkande av händelser och utveckling av begreppsmässiga strukturer. Innefattar vanligen bearbetning av intervjumaterial enligt en kvalitativ metod, exempelvis innehållsanalys, fenomenologi eller grounded theory.

Kvasiexperimentell studie
(eng. *quasi-experimental study*)
▶Experimentell studie utan randomisering.

Kvotskala
(eng. *ratio scale*)
Typ av ▶mätskala. Variabler på kvotskalenivå är sådana där skillnaderna mellan skalsteg i likhet med en ▶intervallskala har samma innebörd oavsett var på skalan man befinner sig, men som också har en absolut nollpunkt (exempelvis vikt, längd). Detta medför att man kan göra meningsfulla jämförelser mellan uppmätta värden, såsom att ett föremål på fyra kilo väger dubbelt så mycket som ett på två kilo.

Longitudinell studie
(eng. *longitudinal study*)
Undersökning där deltagarna följs över tid, d.v.s. mäts vid flera tillfällen. ▶Tvärsnittsstudie.

Marginalanalys
(eng. *marginal analysis*)
En undersökning av konsekvenserna i kostnader och ▶effekter vid små förändringar av en variabel.

Marginalkostnader
(eng. *marginal costs*)
Skillnaden i kostnad för en intervention när man gör små förändringar i dess effektivitet.

Matchning
(eng. *matching*)
Åtgärder för att åstadkomma maximal jämförbarhet mellan grupper i ▶kontrollerade studier där ▶randomisering inte använts, samt i ▶fall-kontrollstudier. För varje fall utses en eller flera matchade kontroller som liknar fallen i avseenden som antas vara viktiga för utfallet.

Medelvärde
(eng. *mean, arithmetic mean*)
Aritmetiskt medelvärde, genomsnitt. Summan av en mängd värden dividerat med antalet observationer.

Medianvärde
(eng. *median*)
Det värde i en rangordnad talserie som delar serien i två lika stora delar. Om serien har ett jämnt antal observationer är medianen medelvärdet mellan de två tal som ligger i mitten. ▶Fraktiler.

Mediatorvariabel
(eng. *mediating variable*)
En faktor som ligger mellan den ▶oberoende och ▶beroende variabeln i en orsakskedja och som helt eller delvis förklarar varför exempelvis en behandlingseffekt äger rum. I till exempel föräldrastödsprogram antas ett förändrat beteende hos föräldrar vara den mediatorvariabel som ger en effekt på barnens beteende. ▶Moderatorvariabel.

Metaanalys
(eng. *meta-analysis*)
En statistisk metod för att sammanväga (poola) resultat från flera enskilda studier som rör en och samma fråga, vanligtvis frågan om en interventions effekter på ett visst utfall. Genom att beakta resultat från flera studier samtidigt får man en bättre möjlighet att uttala sig om interventionens effekter. Ofta använd metod i ▶systematiska översikter.

Minnesbias
(eng. *recall bias*)
Snedvridning av resultat som beror på att individer minns händelser fel eller i olika utsträckning. Denna form av bias är ett hot mot ▶validiteten i ▶retrospektiva studier. Exempelvis om forskare i en undersökning av samband mellan läkemedel och missbildningar intervjuar kvinnor om läkemedel de använt under graviditeten. Kvinnor som fött barn med missbildning har sannolikt övervägt tänkbara orsaker och minns därför bättre läkemedel de tagit.

Modellutvärdering
(eng. *efficacy trial*)
Utvärdering av en ▶intervention när den ges under optimala forskningsförhållanden där forskaren garanterar behandlingstrohet. ▶Verksamhetsutvärdering.

Moderatorvariabel
(eng. *moderator variable*)
En moderatorvariabel påverkar (modererar) riktningen eller styrkan hos ett samband mellan en ▶oberoende variabel och en ▶beroendevariabel. Om exempelvis män och kvinnor har olika utfall efter en behandling är kön en moderatorvariabel av behandlingseffekten. ▶Mediatorvariabel.

Mothypotes
▶Alternativ hypotes.

Multi-center-studie
(eng. *multi-site study, multi-center study*)
En studie som utförs vid flera centra (kliniker, socialkontor etc.) för att man inom rimlig tid ska kunna inkludera ett tillräckligt stort antal deltagare. En annan fördel är att resultatens generaliserbarhet ökar. Kallas även multi-site-studie.

Multi-informant-studie
(eng. *multi-informant study*)
En studie som använder flera parallella uppgiftslämnare för att mäta ett utfall, exempelvis när man för att mäta ungdomars beteende gör mätningar även med deras föräldrar och lärare.

Multi-metod-studie
(eng. *multi-method study*)
En studie som använder flera typer av datakällor (metodtriangulering), exempelvis registerdata och självrapportering från deltagare. Till skillnad från en studie där man använder både kvalitativ och kvantitativ forskningsmetodik (eng. mixed-method study), avser termen multi-metod-studie användning av olika metoder inom ett och samma paradigm.

Mätskala
(eng. *level of measurement*)
System för klassificering av ▶variabler som baseras på deras underliggande egenskaper. En variabels mätskala (även benämnd datanivå) bestämmer vilken typ av statistik som kan användas när man analyserar den. Det finns fyra mätskalor med olika typ och mängd av information, ▶nominalskala, ▶ordinalskala, ▶intervallskala och ▶kvotskala.

Narrativ översikt
(eng. *narrative review*)
En narrativ (berättande) översikt ger ofta en övergripande beskrivning av ett ämne snarare än att söka besvara en specifik fråga, som hur effektiv en viss intervention är för ett visst tillstånd (problem). Narrativa översikter redovisar sällan hur sökningen efter litteratur gick till eller vilka kriterier som använts för att välja ut det som inkluderats. En narrativ översikt skiljer sig från en ▶systematisk översikt med narrativ syntes. I det senare fallet används en narrativ syntes för att summera resultat från de ingående studierna då statistisk sammanvägning (▶metaanalys) inte är möjlig eller lämplig, samtidigt som översikten uppfyller den systematiska översiktens krav på systematik och transparens.

Nollhypotes (H_0)
(eng. *null hypothesis*)
Vid ▶statistisk hypotesprövning ett i förväg uppställt antagande om att det inte finns något samband mellan de fenomen som studeras. Nollhypotesen för en ▶effektstudie är att det inte finns någon skillnad mellan gruppernas utfall vid uppföljningsmätningen. ▶Alternativ hypotes.

Nominalskala
(eng. *nominal scale*)
Typ av ▶mätskala. Variabler på nominalskalenivå (även kallade kategorivariabler) är sådana som inte kan rangordnas och vars värden återspeglar egenskaper eller namn, exempelvis hårfärg, kön eller religion.

Normalfördelning (eng. *normal distribution, Gaussian distribution, bell curve*)
En teoretisk fördelning som normalt sett används som bas för analyser av kontinuerliga data (t.ex. vikt i kilo). En normalfördelad variabel antar ofta värden som ligger nära medelvärdet och sällan värden som har stor avvikelse. Därför ser normalfördelningen ut som en kulle, d.v.s. en symmetrisk klockformad kurva där de båda ”svansarna” är lika långa. Kan beskrivas med centralmåttet ▶medelvärde och spridningsmåttet ▶standardavvikelse.

Normering
Normering innebär att man administrerar ett ▶standardiserat instrument till ett ▶urval från en ▶population för att erhålla ett referensmaterial. Detta kan sedan användas för jämförelse i kommande kliniskt arbete eller forskning för att ge en uppfattning om vad ett erhållet mätvärde från en individ eller grupp innebär.

Number needed to treat (NNT)
En kliniskt relevant ▶effektstorlek som beskriver hur många individer som måste få en intervention för att man ska förebygga ett enda oönskat utfall. Om exempelvis fem ungdomar måste genomgå en viss behandling mot kriminalitet innan en upphör med brottslighet blir NNT fem.

Oberoende variabel (eng. *independent variable, explanatory variable; predictor variable*)
Avser en ▶variabel som antas kunna påverka värdet på en annan variabel. Inom ramen för en ▶effektutvärdering är den primära oberoende variabeln deltagarnas grupptillhörighet (interventions- eller kontrollgrupp). ▶Beroendevariabel.

Observationsstudier (eng. *observational studies*)
Sammanfattande term för de forskningsdesigner som används inom epidemiologisk forskning (▶kohortstudier, ▶fall-kontrollstudier och ▶tvärsnittsstudier), ofta i syfte att identifiera tänkbara orsaker till sjukdom i stora deltagarmaterial. Till skillnad från en ▶experimentell studie där forskaren medvetet introducerar en intervention för att avläsa dess effekter, vidtar forskaren i observationsstudier ingen aktiv åtgärd. I stället observeras naturligt förekommande samband mellan förmodade sjukdomsorsaker och förekomsten av sjukdom. I observationsstudier finns vanligtvis en större risk för ▶selektionsbias än i ▶kontrollerade studier.

Oddskvot
(eng. *odds ratio, OR*)
Kvoten mellan två ▶oddstal. Till exempel oddstalet att individer i en behandlingsgrupp avlidit efter fem år dividerat med motsvarande oddstal för ▶kontrollgruppen. En oddskvot på 1 innebär att det inte finns någon skillnad mellan grupperna. ▶Riskkvot.

Oddstal, odds
(eng. *odds*)
Antalet fall av en händelse dividerat med antalet fall där händelsen inte inträffat. Om exempelvis 20 av 100 personer avlidit fem år efter en behandling så är oddset för dödsfall 20/80 = 0.25 eller 1:4. ▶Risk.

Operationalisering
(eng. *operationalization*)
Att definiera ett begrepp i mätbara termer. Kriminalitet kan exempelvis operationaliseras som antal registrerade domar.

Ordinalskala
(eng. *ordinal scale*)
Typ av ▶mätskala. Variabler på ordinalskalenivå är sådana där värdena kan rangordnas. Till skillnad från variabler på ▶intervallskala och ▶kvotskala vet man dock inte vad avståndet mellan två skalsteg innebär. Om man exempelvis har resultaten från en löpartävling på ordinalskalenivå vet man löparnas placering i resultatlistan (första, andra etc.) men ingenting om deras sluttider.

Orsaksvariabel
▶Oberoende variabel.

Outlier
▶Extremvärde.

Ovidkommande variabel
(eng. *confounding factor; extraneous variable*)
Inom en effektutvärdering en variabel som samvarierar både med den ▶oberoende variabeln och ▶beroendevariabeln och som medför en risk att interventionens effekter överskattas eller underskattas. Ovidkommande variabler undviks bäst med ▶randomisering som teoretiskt sett likställer grupperna vid studiens början, även på alla de variabler som inte mäts inom studien. Man kan också kontrollera för ovidkommande variabler med hjälp av statistik, men då endast de variabler som man mätt inom ramen för studien.

Parameter
(eng. *parameter*)
Inom statistiken ett värde (t.ex. ▶medelvärde) som avser en hel ▶population. En parameter kan skattas utifrån ett värde som erhållits från ett ▶urval av populationen (▶estimat).

Percentil
(eng. *percentile*)
▶Fraktiler.

Placebokontroll
(eng. *placebo-control*)
En overksam ▶intervention som administreras till en grupp deltagare i en ▶kontrollerad studie. Den används för att kontrollera för den effekt som kan uppstå av att deltagare tror att de får en verklig behandling (placeboeffekt). Placebokontroll används regelbundet i klinisk-medicinska utvärderingar (”sockerpiller”) men är svårare att tillämpa i studier av psykologiska och sociala interventioner.

Population
(eng. *population*)
En grupp personer som har något gemensamt, till exempel alla personer i Sverige eller alla arbetslösa personer i en kommun. Vid stora populationer görs undersökningar sällan som ▶totalundersökningar utan på ett ▶urval från populationen, till exempel personer med en viss diagnos vid en viss mottagning som uppfyller ▶inklusions- och ▶exklusionskriterier, och som gett ▶informerat samtycke att delta i undersökningen. En sådan urvalsgrupp kan också kallas en studiepopulation eller stickprov.

Post hoc-analyser
(eng. *post-hoc analyses*)
Statistiska analyser som inte var planerade vid studiens start.

Power
(eng. *statistical power*)
▶Statistisk styrka.

Primärstudie
(eng. *primary study*)
Rapport från en empirisk studie där individdata samlats in. Primärstudier särskiljs från ▶sekundärstudier som innebär analys och redovisning av resultat från tidigare genomförda primärstudier (t.ex. ▶systematisk översikt).

Produktivitetskostnader
(eng. *productivity costs*)
Kostnader förknippade med utebliven eller nedsatt arbetsförmåga eller möjlighet att delta i fritidsaktiviteter på grund av sjukdom eller dödsfall.

Programtrohet
▶Behandlingstrohet.

Prospektiv studie
(eng. *prospective study*)
En studie som följer människor framåt i tiden. Kontrollerade ▶experimentella studier är alltid prospektiva. ▶Kohortstudier är antingen prospektiva eller retrospektiva, medan ▶fall-kontroll-studier oftast är retrospektiva. ▶Retrospektiv studie.

Publikationsbias
(eng. *publication bias*)

Otillförlitlighet i resultaten från en ▶systematisk översikt som innebär en risk att den sammanvägda effekten av en intervention överskattas. Uppstår på grund av att forskare och tidskriftsredaktörer tenderar att publicera relativt fler studier med signifikanta, positiva resultat än studier med nollresultat eller med negativa resultat. En liknande snedvridning kan ske på utfallsnivå och kallas då ▶rapporteringsbias.

P-värde
(eng. *p-value*)

Probabilitetsvärde, uttryck för ▶signifikansnivå. Beskriver sannolikheten för att man av en slump skulle få det erhållna resultatet eller ett mer extremt resultat om ▶nollhypotesen vore sann. P-värdet ger ingen information om ▶effektstorleken.

Randomiserad kontrollerad studie
(eng. *randomized controlled trial, RCT*)

▶Kontrollerad studie där deltagare fördelas slumpmässigt till ▶interventionsgrupp och ▶kontrollgrupp. Randomiserade kontrollerade studier ger de mest tillförlitliga resultaten avseende interventioners effekter eftersom man i teorin likställer grupperna på kända och okända ▶ovidkommande variabler (eng. *confounders*) vid studiens start och därmed minskar risken för ▶selektionsbias.

Randomisering
(eng. *randomization, random assignment, random allocation*)

Slumpmässig fördelning (allokering) av deltagare till grupper i en ▶randomiserad kontrollerad studie. Allokeringsordningen bestäms med hjälp av exempelvis en slumptalstabell eller slumptalsgenerator. För att randomiseringen ska tjäna sitt syfte att likställa grupperna vid studiens början bör allokeringsordningen vara dold (omöjlig att förutse) för forskare och deltagare. Termen ska ej förväxlas med slumpmässigt urval, se ▶slumpmässig.

Rapporteringsbias
(eng. *reporting bias*)

Otillförlitlighet gällande interventioners effekter på enskilda utfall. Uppstår när forskare i sin rapportering utelämnar utfallsmått som inte påverkats av interventionen alternativt lägger till utfallsmått (där positiva effekter påvisats) som inte var tänkta att inkluderas från början. En liknande snedvridning kan ske på studienivå. ▶Publikationsbias.

RCT

▶Randomiserad kontrollerad studie.

Reaktivitet
(eng. *reactivity*)
Avser att deltagare i en studie ändrar sitt beteende eftersom de vet att de är observerade (▶Hawthorne-effekt) eller för att försöka motsvara förväntningar som försöksledaren med små signaler kan ha förmedlat (▶försöksledareffekt). Reaktivitet är ett hot mot en studies ▶interna validitet och kontrolleras vanligtvis genom ▶blindning.

Registerstudie
(eng. *registry study*)
▶Retrospektiv eller kombinerad retro- och ▶prospektiv undersökning som baseras på data från offentliga register.

Regression, regressionsanalys
(eng. *regression analysis*)
En grupp av statistiska analyser som används för att undersöka sambandet mellan en beroendevariabel och en eller flera förklarande variabler. Den beroende variabeln benämns ofta utfalls- eller responsvariabel. Den förklarande variabeln benämns ofta exponeringsvariabel, kovariat eller prediktor. Vanliga varianter är linjär och logistisk regressionsanalys.

Relativ risk
▶Riskkvot.

Reliabilitet
(eng. *reliability*)
Den grad med vilken man kan upprepa ett resultat som erhållits med en mätprocedur, t.ex. ett ▶standardiserat instrument. Reliabilitet uttrycks som en ▶korrelation med värden mellan 0.00 och 1.00, där högre värden anger högre grad av stabilitet hos mätproceduren. Av reliabilitet följer inte att testet mäter det som det är avsett att mäta, har ▶validitet. Om reliabiliteten brister kan ett test eller en mätning dock inte heller anses valid. ▶Test-retest-reliabilitet, ▶interbedömarreliabilitet, ▶intern konsistens.

Residual, residualvarians
(eng. *residual variance*)
Kallas även fel eller felvarians. Den variation i utfallsvariabeln som inte förklaras av en regressionsmodell. Residualen är skillnaden mellan en dataobservations faktiska värde på utfallsvariabeln och modellens predicerade värde för den observationen.

Resultatvariabel
▶Utfallsmått.

Resurser
(eng. *resources*)
Alla poster inom ekonomin som kan användas för att producera och distribuera varor och tjänster.

Retrospektiv studie
(eng. *retrospective study*)
Undersökning som är tillbakablickande, d.v.s. utnyttjar data som insamlats eller händelser som inträffat innan undersökningen startas. Ett exempel är ▶fall-kontrollstudier och ibland även ▶kohortstudier. ▶Prospektiv studie.

Risk för bias (eng. *risk of bias*)
Inom ramen för ▶systematiska översikter av företrädesvis ▶interventioners effekter bedöms risken för olika typer av ▶bias hos de ingående ▶primärstudierna. Bedömningen avser risken att det erhållna resultatet för ett utfallsmått från primärstudierna (och därmed från översikten) är otillförlitligt. Enligt metodik från ▶Cochrane Collaboration bedöms risk för bland annat ▶selektionsbias i ▶randomiserade kontrollerade studier genom att man granskar om fördelningsordningen genererats slumpmässigt och om den varit dold (omöjlig att påverka) för forskare och deltagare. För varje utfallsmått bedöms risken för varje typ av bias som antingen låg, oklar eller hög.

Riskkvot (eng. *risk ratio, relative risk*)
Annan benämning är relativ risk. En ▶effektstorlek som utgörs av kvoten mellan ▶risktalen i två undersökta grupper. I en ▶interventionsstudie är detta kvoten mellan risken i ▶interventionsgruppen och risken i ▶kontrollgruppen. En riskkvot på 1 innebär att ingen skillnad finns mellan grupperna. En riskkvot under 1 innebär att interventionen effektivt minskat risken för det oönskade utfallet.

Risktal, risk (eng. *risk*)
Sannolikheten att en negativ händelse ska inträffa i en grupp. Den beräknas som antalet personer som drabbades av händelsen dividerat med totala antalet personer i gruppen. ▶Oddstal.

Rörliga kostnader (eng. *variable costs*)
Kostnader som varierar med den mängd som produceras.

Samband (statistiskt)
▶Korrelation.

Sampel
▶Urval.

Statens beredning för medicinsk utvärdering, SBU (eng. *Swedish Council on Health Technology Assessment*)
Den myndighet som har regeringens uppdrag att utvärdera olika metoder i vården ur ett samlat medicinskt, ekonomiskt, etiskt och socialt perspektiv. Arbetet bedrivs huvudsakligen genom produktion av ▶systematiska översikter.

Sekundärstudie (eng. *secondary study*)
En studie där man inte samlar in egna data utan i stället redovisar resultat från redan genomförda ▶primärstudier (t.ex. ▶systematisk översikt).

Selektionsbias
(eng. *selection bias*)

Otillförlitlighet i resultat från en ▶kontrollerad studie som kan uppstå när de studerade grupperna skiljer sig från början vad gäller prognos eller mottaglighet för en behandling. Framför allt ▶randomisering men även ▶matchning minskar risken för selektionsbias. Selektionsbias är ett hot mot studiens ▶interna validitet.

Sensitivitetsanalys
(eng. *sensitivity analysis*)

Sensitivitetsanalys innebär att man undersöker hur känsliga resultaten från en studie eller ▶systematisk översikt är för de metodval och antaganden som gjorts i samband med den statistiska analysen. Ett exempel är när man använder flera olika metoder för ▶imputering av saknade värden i ett dataset, och därefter jämför resultaten från statistiska analyser som gjorts under dessa olika förhållanden. Om resultaten är lika mellan olika imputeringsmetoder är det ett tecken på att imputeringsresultaten är tillförlitliga.

Signifikansnivå
(eng. *significance level, alpha level*)

Ett statistiskt begrepp inom hypotesprövning. Signifikansnivån anger sannolikheten för att man av en slump skulle få det erhållna resultatet eller ett mer extremt resultat om ▶nollhypotesen vore sann. Signifikansnivån uttrycks som ett p-värde där p står för probabilitet (sannolikhet). Ofta väljs $p = .05$ som den högsta accepterade risken att felaktigt förkasta nollhypotesen. Risken för ▶typ I-fel är då fem procent. Andra vanliga signifikansnivåer (alfanivåer) är $p < .01$ och $p < .001$. Det p-värde som erhålls vid den statistiska analysen avgör om och med vilken grad av säkerhet man kan förkasta nollhypotesen.

Skevhet i datafördelning
(eng. *skew*)

Skevhet (även kallad snedhet) och ▶toppighet utgör två beskrivande egenskaper hos en sannolikhetsfördelning. Till skillnad från en ▶normalfördelning är de båda ”svansarna” i en skev fördelning olika långa, med en mer långsmal svans på en av sidorna. I en positivt skev fördelning är svansen mer långsmal på höger sida, och i en negativt skev fördelning är svansen mer långsmal på vänster sida.

Skuggpris
(eng. *shadow price*)

Samhällets ▶alternativkostnad för ett resultat.

Slumpmässig
(eng. *random*)

Styrd av slumpen. Slumpmässighet är centralt vid fördelningen av deltagare till grupper i ▶randomiserade kontrollerade studier samt vid ▶urval av undersökningsdeltagare i tvärsnittsstudier av en population. I det första fallet reducerar man risken för ▶selektionsbias och ökar därmed chanserna till en god ▶intern validitet. I det andra fallet erhåller man ett urval som är representativt för populationen som helhet, vilket ökar chanserna till en god ▶extern validitet.

Standardavvikelse
(eng. *standard deviation*)

Ett spridningsmått för en sannolikhetsfördelning; hur utspridd fördelningen är kring ett ▶medelvärde. En liten standardavvikelse (SD) innebär att de flesta mätvärden ligger nära medelvärdet. Standardavvikelsen utgör kvadratroten ur ▶variansen. Som spridningsmått är SD intuitivt mer förståelig än variansen eftersom den anges på samma skala som medelvärdet.

Standardbehandling

▶Traditionell behandling.

Standardiserat instrument
(eng. *standardized instrument*)

Nomenklaturen kring bedömningsmetoder, instrument och standardisering är ännu inte enhetlig. En standardiserad bedömningsmetod kan dock sägas bestå av ett instrument med tillhörande manual. Instrumentet är ett formulär med i förväg fastställda frågor och svarsalternativ i form av numerär eller verbal skala, där svaren kan summeras eller räknas fram enligt en bestämd formel. Instrument kan vara tester, intervjuformulär, skattningsformulär eller observationsformulär. Termen ”standardiserad” används med olika betydelser där den mest grundläggande endast innebär att innehållet är fastställt och utprövat. Enligt Standardiseringsinstitutet innebär standardisering att bedömningsmetoden är baserad på detaljerade regler och specifikationer, inklusive administrativa riktlinjer utarbetade av utvecklaren med syfte att uppnå ett enhetligt, bestående bedömningsförfarande. Ett standardiserat instrument kan också vara ▶normerat.

Statistisk hypotesprövning
(eng. *statistical hypothesis testing*)
Användande av statistiska test för att undersöka om ▶nollhypotesen ska antas (när ingen skillnad föreligger mellan de undersökta grupperna) eller förkastas (när skillnad föreligger varvid den ▶alternativa hypotesen antas).

Statistisk regression
(eng. *statistical regression to the mean*)
Inträffar när deltagare i en studie har extrema värden vid en förmätning och dessa blir mindre extrema vid en uppföljningsmätning enbart för att extrema värden tenderar att närma sig medelvärdet när en mätning upprepas. I en behandlingsstudie där problemnivån vanligen är hög när deltagare rekryteras kan man alltså förvänta sig en lägre problemnivå vid uppföljning enbart på grund av att datafördelningen statistiskt har dragit sig mot medelvärdet. Detta är ett hot mot den ▶interna validiteten i ▶interventionsstudier och kontrolleras genom inklusion av en obehandlad ▶kontrollgrupp i ▶studiedesignen.

Statistisk styrka
(eng. *statistical power*)
Statistisk sensitivitet; sannolikheten för att en ▶effektutvärdering ska kunna påvisa en effekt av interventionen givet att en effekt existerar. Om styrkevärdet är 0,80 innebär det att forskaren har 80 procents chans att kunna påvisa en befintlig skillnad. Styrkan anger sannolikheten för att man ska undgå att göra ett ▶typ II-fel, d.v.s. att betrakta en verklig skillnad som icke-existerande. Vid planeringen av en studie görs en beräkning som baseras på styrkevärdet, önskad ▶signifikansnivå och den förväntade ▶effektstorleken. Beräkningen visar hur många deltagare studien behöver för att man ska hitta en effekt givet att en sådan finns.

Statistisk validitet
(eng. *statistical conclusion validity*)
Statistiska aspekter av en ▶effektstudie som påverkar slutsatserna om interventionens effekter. Ett hot mot den statistiska validiteten är låg ▶statistisk styrka.

Studie
(eng. *study, trial*)
Allmän benämning på en vetenskaplig undersökning.

Studiedesign
(eng. *study design*)
Det studieupplägg som används för att besvara en forskningsfråga. Exempel på studiedesigner är ▶randomiserade kontrollerade studier och ▶observationsstudier.

Studieprotokoll ▶Forskningsplan.

Störfaktor ▶Ovidkommande variabel.

Systematisk kartläggning (eng. *scoping study*)
En systematisk kartläggning följer i princip samma metodologi som en ▶systematisk översikt, men frågeställningen är ofta bredare formulerad och omfattar inte någon värderande sammanvägning av resultat. En systematisk kartläggning ger kunskap om vad det finns för forskning om ett i förväg avgränsat ämnesområde. Genom att klassificera forskningen med avseende på bibliografiska uppgifter och huvudsakligt innehåll kan systematiska kartläggningar tydliggöra kunskapsluckor. Resultaten används därför ofta som utgångspunkt för framtida studier, såväl systematiska översikter som ▶primärstudier.

Systematisk översikt (eng. *systematic review*)
En systematisk översikt är en strukturerad metod som används för att identifiera, välja ut, bedöma och sammanfatta forskning avseende en tydlig och avgränsad fråga. Frågan avser vanligtvis (men inte alltid) effekter av interventioner och kan sammanfattas med akronymen PIKU – för Population a, har Intervention b större effekt än Kontrollintervention c på Utfall d? Ambitionen är att så systematiskt och transparent som möjligt väga samman forskningsresultat från alla kända och relevanta ▶primärstudier som håller acceptabel vetenskaplig kvalitet. De uttömmande anspråken medför att den systematiska översikten är ett levande dokument som ska revideras regelbundet. Arbetet genomförs i enlighet med ett protokoll där samtliga arbetsmoment har specificerats i förväg. ▶Cochrane Collaboration har haft stor betydelse för metodiken för systematiskt översiktsarbete. PRISMA statement tillhandahåller riktlinjer för hur systematiska översikter ska rapporteras och AMSTAR är ett instrument för bedömning av deras metodologiska kvalitet.

Systematiskt fel ▶Bias.

Test-retest-reliabilitet (eng. *test-retest reliability*)
Mått på replikerbarheten (stabiliteten) hos en mätprocedur, t.ex. ett ▶standardiserat instrument. Uppskattas genom att man administrerar instrumentet till samma individer vid upprepade tillfällen, under liknande betingelser. Om det finns en hög ▶korrelation mellan mätningarna har instrumentet hög test-retest-reliabilitet. ▶Reliabilitet.

Toppighet i datafördelning (eng. *kurtosis*)
Toppighet och ▶skevhet i datafördelning utgör två beskrivande egenskaper hos en sannolikhetsfördelning. Graden av toppighet (kurtos) säger något om sannolikheten för att fördelningens värden är extrema, d.v.s. ligger långt från medelvärdet. En normalfördelad (mesokurtosisk) variabel med kurtos 3 uppvisar en klockformad fördelning med få extremvärden, medan en smalare och toppigare (leptokurtosisk) kurva med kurtos större än 3 har tjocka svansar och fler extremvärden. En plattare fördelning med smala eller inga svansar kallas platykurtosisk.

Totala kostnader (eng. *total costs*)
Summan av alla kostnader för en ▶intervention eller ett problem.

Totalundersökning (eng. *census*)
Statistisk undersökning som innefattar en hel ▶population och inte enbart ett ▶urval.

Traditionell behandling (eng. *Treatment-As-Usual; TAU*)
Ofta använd jämförelsegrupp inom ▶effektutvärdering som utgörs av den normalt förekommande insatsen för en viss målgrupp vid ett visst tillfälle. Även kallad standardbehandling och konventionell behandling.

Treatment-on-the-treated-analys (TOT)
Resultatbearbetning i en ▶effektutvärdering som endast omfattar deltagare som fullföljt interventionen eller behandlingen i tillräckligt hög grad. TOT rekommenderas vanligen inte eftersom de individer som inte fullföljt behandlingen kan förväntas skilja sig från dem som fullföljt. Bortfallet kan också se olika ut i grupperna; en intervention kan exempelvis ha varit bättre på att behålla de mest svårbehandlade, vilket kommer att påverka resultatet. Istället rekommenderas vanligen ▶intention to treat-analys vid effektutvärderingar.

Trimmat medelvärde
(eng. *trimmed mean*)
En robust typ av ▶medelvärde som räknas fram efter att en bestämd andel av de högsta och lägsta värdena uteslutits. Borttagna värden kan vara ▶outliers men behöver inte vara det.

Tvärsnittsstudie
(eng. *cross-sectional study*)
Undersökning där deltagarna mäts vid ett enda tillfälle. Används exempelvis för att mäta förekomst av ett tillstånd eller problem. ▶Longitudinell studie.

Typ I-fel
(eng. *Type 1-error*)
Felaktigt förkastande av ▶nollhypotesen när den är sann, såsom slutsatsen att en intervention har varit effektiv när den inte har varit det (även kallat alfafel). Risken för typ I-fel motsvarar ▶signifikansnivån för ett statistiskt test, så när $p < .05$ finns 5 procents risk för ett typ I-fel. Eftersom risken för typ-I-fel ökar med antalet genomförda test väljer man ofta att korrigera signifikansnivån nedåt med så kallad ▶Bonferronikorrektion.

Typ II-fel
(eng. *Type II-error*)
Felaktigt accepterande av nollhypotesen när den inte är sann, såsom slutsatsen att en intervention inte haft effekt när den har haft det (även kallat beta-fel). Risken för typ II-fel är direkt kopplad till studiens ▶statistiska styrka.

Urval
(eng. *sample*)
Stickprov. Eftersom det i allmänhet är omöjligt att inkludera samtliga individer i en ▶population i en studie, görs oftast ett urval av undersökningspersoner. Idealt ska urvalet vara ▶slumpmässigt genomfört, vilket gör det representativt för populationen som helhet.

Urvalsbias
▶Selektionsbias.

Utfallsmått, utfallsvariabel
(eng. *outcome variable*)
Aspekter av en persons kliniska och funktionella status som mäts inom ramen för en ▶effektutvärdering och som fungerar som kriterium för att värdera interventionens effekter, exempelvis hur stor andel i vardera gruppen som inom viss tid återfallit i brottslighet.

Validitet, experimentell
(eng. *validity*)
Inom ramen för en ▶effektutvärdering avser validitet giltigheten, trovärdigheten hos olika aspekter av studien. ▶Intern validitet, ▶Extern validitet, ▶Statistisk validitet.

Validitet, testteoretisk (eng. *validity*)
Inom testteori avser validitet allmänt sett empiriskt stöd för att test och mätningar mäter det som man avsett att mäta. ▶Begreppsvaliditet, ▶innehållsvaliditet, ▶kriterievaliditet.

Variabel (eng. *variable*)
En faktor som (till skillnad från en konstant) kan anta olika värden. ▶Oberoende variabel, ▶beroendevariabel.

Varians (eng. *variance*)
Ett spridningsmått för en sannolikhetsfördelning; hur utspridd fördelningen är kring ett ▶medelvärde. En liten varians innebär att de flesta mätvärden ligger nära medelvärdet. ▶Standardavvikelse.

Variansanalys (eng. *analysis of variance, ANOVA*)
En grupp av statistiska analyser som används för att undersöka om medelvärden mellan två eller flera grupper av individer skiljer sig signifikant åt. Skillnaderna mellan gruppernas medelvärden (mellangruppsvariansen) ställs i relation till skillnaderna mellan individer inom de olika grupperna (inomgruppsvariansen). Vanliga varianter är envägs-ANOVA, faktoriell ANOVA och ANOVA för upprepad mätning.

Variation (eng. *variation*)
Spridningen av mätvärden i en datamängd. Det finns flera spridningsmått, exempelvis ▶varians, ▶standardavvikelse, ▶konfidensintervall, ▶fraktiler.

Verksamhetsutvärdering (eng. *effectiveness trial*)
Utvärdering av en intervention när den ges i ordinarie verksamhet. Kontrasteras ofta mot ▶modellutvärdering.

Väntelista (eng. *waiting-list control group*)
En typ av jämförelsegrupp i en ▶kontrollerad studie. Deltagare som ställs på väntelista får den studerade ▶interventionen vid ett senare tillfälle, vanligtvis när den första uppföljningsmätningen genomförts. Man får då en jämförelse mellan en obehandlad och en behandlad grupp samtidigt som alla deltagare erhåller den studerade insatsen.

Z-poäng (eng. *Z-score*)
Standardiserade värden som anger hur många ▶standardavvikelser från en datamängds ▶medelvärde ett observerat värde befinner sig. En z-poäng över 3 brukar beteckna en ▶outlier. ▶Box-plots används med fördel för identifikation av univariata outliers.

Årliga kostnader (eng. *annual costs*)
Kostnader som uppstår varje år.

Återkommande kostnader (eng. *recurrent costs*)
Värdet av resurser som måste köpas minst en gång per år.

Åtgärd
▶Intervention.

Överförbarhet (eng. *relevance, directness*)
Överförbarhet är ett begrepp som både avser enskilda studiers ▶externa validitet och som utgör en bedömningsgrund inom evidensgraderingssystemet ▶GRADE.

Överföring (eng. *transfer*)
En betalning som görs till en enskild individ som inte utför någon tjänst i utbyte mot betalningen, till exempel sociala avgifter eller arbetslöshetsersättning.

Register