AKF Rapport

Olaf Rieper og Hanne Foss Hansen

Metodedebatten om evidens

Metodedebatten om evidens

af

Olaf Rieper
Hanne Foss Hansen

AKF Forlaget
Oktober 2007

AKF's publikationer forhandles gennem
boghandelen og AKF Forlaget,
Nyropsgade 37, 1602 København V
Telefon: 43333400 eller Fax: 43333401
E-mail: akf@akf.dk
Internet http://www.akf.dk

Forlag: AKF Forlaget
Tryk: Litotryk København A/S
Isbn. nr.: 978-87-7509-833-0
I\Forlaget\OR\Metodedebatten om evidens\.jp

Oktober 2007

AKF, Anvendt KommunalForskning
AKF har til formål at gennemføre
og formidle samfundsforskning af
relevans for det offentlige og især for
regioner og kommuner.

Forord

Rapporten udgives fra projektet: »Metaevaluering: Akkumulering af viden i et vidensamfund?«. Projektet er støttet af Forskningsrådet for Samfund og Erhverv under Det Frie Forskningsråd i Danmark og gennemføres i samarbejde mellem Institut for Statskundskab ved Københavns Universitet og AKF. Projektet løber indtil ultimo 2007.

Metoderapporten er den anden rapport fra projektet, der er tidligere udgivet en rapport, der kortlægger evidensbevægelsens udbredelse, organisering og arbejdsform, idet der sættes fokus på »anden ordens« videninstitutioner forstået som organisationer, der sammenfatter resultaterne fra flere enkeltstående undersøgelser.

Denne rapport giver en oversigt over centrale metodiske positioner og diskussioner inden for evidensbevægelsen. Vi har valgt en beskrivende, analytisk vinkel og kommer på den baggrund til sidst med vores egne synspunkter på udviklingen.

Projektet ledes af professor Hanne Foss Hansen, Institut for Statskundskab, Københavns Universitet, i samarbejde med programchef Olaf Rieper, AKF. Yosef Bhatti og Leo Milgrom har som studentermedarbejdere bidraget med analyse af baggrundsmaterialer.

Hanne Foss Hansen
Olaf Rieper

Oktober 2007

Indhold

1 Sammenfatning

Sigtet med denne rapport er at give en oversigt over centrale metodiske positioner og diskussioner inden for det, vi kalder evidensbevægelsen. Begrebet »evidens« er blevet et plusord i offentlig forvaltning: evidensbaseret politik, evidensbaseret praksis, evidensbaseret ledelse, evidensbaseret medicin, pædagogik osv. Evidensbevægelsen er optaget af at sammenfatte viden fra flere enkeltstående undersøgelser og evalueringer. Formålet er at producere og formidle den bedst mulige viden om resultaterne af givne interventioner eller indsatser. Evidensbevægelsen er vokset frem internationalt og nationalt i de sidste 10-15 år. Vi afgrænser os til det, der efter vores mening er det nye i evidensbegrebet, nemlig at der er etableret globale og nationale organisationer og netværk, der er specialiserede i at producere, bestille og formidle forskningsoversigter til beslutningstagere i politik og praksis – og som er tilgængelig for alle. Forskningsoversigter er sammenfatninger af foreliggende forskning og undersøgelser om et givet emne, fx effekter af en bestemt indsats, intervention eller behandling, foretaget på systematiske og gennemskuelige måder. På engelsk betegnes forskningsoversigterne som »systematic reviews«. Vi sætter fokus primært på de store velfærdsområder, sundheds-, social- og uddannelsesområdet.

Vores hovedkonklusion er, at de til tider heftige diskussioner for og imod et smalt henholdsvis bredt evidensbegreb udfolder sig forskelligt på de forskellige sektorområder (sundheds-, uddannelses- og socialområdet). Diskussionerne er i betydelig grad præget af de faglige traditioner og interesser, som kendetegner professionsgrupperne på de forskellige sektorområder. Tidligere fandt metodediskussioner hovedsageligt sted i forsknings-

kredse. Metodediskussionerne er med evidensbevægelsen flyttet uden for forskningens mure til det politiske og fagprofessionelle niveau. De mulige konsekvenser bliver, at uddannelse, socialarbejde og sundhedstjenester forandres og begrundes såvel praktisk som politisk. Forskningsmæssig viden spiller en stadig større rolle for, hvordan vi former samfundet, herunder den offentlige sektor. Og evidensbevægelsen synes at kunne udgøre et vigtigt bidrag hertil. Faren er, at evidensbasen defineres smalt, idet den kun bygger på lodtrækningsforsøg og kvantitative analyser. Vi mener, at evidensbevægelsen bør organiseres med afsæt i en bred tilgang til forskningsmetoder.

En del af disse organisationer og netværk, der producerer og formidler forskningsoversigter, har bestemte aktører som målgruppe i flere lande. På sundhedsområdet er der fx Cochrane-samarbejdet, og på social- og arbejdsmarkedsområdet samt det kriminologiske område er der Campell-samarbejdet). Andre organisationer har nationale målgrupper, fx det engelske »Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI)«. I en tidligere rapport har vi beskrevet disse organisationer med fokus på Europa (Bhatti, Hansen og Rieper 2006).

Disse organisationers produkter: forskningsoversigterne, har en stor potentiel indflydelse på, hvad der hos målgrupperne – dvs. beslutningstagere på flere niveauer – godtages som troværdig viden. Og dermed er de vigtige skabere af viden, der betragtes som troværdig og legitim for politik og praksis. De afgrænser simpelthen, hvad der kan betragtes som »gyldig« viden. Netop derfor er debatten vigtig i forhold til, hvordan forskningsoversigter frembringes. Det er m.a.o. en debat om, hvilke metoder der skal anvendes, når man udarbejder forskningsoversigter. Det er især ét element i produktionen af forskningsoversigter, der giver anledning til debat. Og det er den kvalitetsvurdering, der gennemføres af primærstudier for at afgøre, om de skal medtages eller ej i en given forskningsoversigt. Der er også metodedebatter på andre punkter, fx omkring metoder til at syntetisere (sammenfatte) resultater fra flere primærstudier. Den debat kommer vi også ind på, men fokus er på kvalitetsvurderingen af primærstudier.

I denne rapport beskriver vi først en af de mest udbredte tankegange i denne vurdering, som benævnes »evidenshierarkiet«. I toppen af hierarkiet er, hvad der benævnes som den gyldne standard for forskningsdesign, det

randomiserede, kontrollerede eksperiment (randomized controlled trials – RCT). Længere nede ad stigen kommer andre design, fx forløbsundersøgelser og endnu længere nede casestudier. Forestillingen er her, at forskningsdesign kan rangordnes, således at nogle design, hvis de gennemføres optimalt, vil give mere troværdige resultater end andre design. Vi beskriver de forskellige design i et typisk evidenshierarki og giver eksempler på undersøgelser baseret på de forskellige design.

For det andet har vi gennemgået retningslinjer og håndbøger fra 10 organisationer i USA og Europa, som producerer evidens. På det grundlag har vi beskrevet, hvordan disse organisationer selv mener, at primærstudier skal vurderes. Det viser sig, at 6 af de 10 organisationer i deres egne retningslinjer angiver, at de arbejder ud fra evidenshierarkiets logik. (Herudover henviser to organisationer til disse organisationers retningslinjer.) De to resterende organisationer angiver, at de ikke tager udgangspunkt i rangorden af design. De skriver bl.a., at deres forskningsoversigter ikke kun drejer sig om effekter af indsatser, men også om implementering, brugeropfattelser etc., hvor andre design end RCT er optimale. De åbner for en bredere videnbase end forskning og medtager således også praksisviden fra fagprofesionelle og brugerforankret viden. Ligesom de integrerer forskning baseret på forskellige design i samme forskningsoversigt. Der er imidlertid også brug for at kvalitetsvurdere primærstudier efter, de er blevet udvalgt på grundlag af deres design. Et RCT kan jo som andre design være gennemført mere eller mindre godt. Her viser det sig, at organisationernes retningslinjer peger på, at de organisationer, der betoner evidenshierarkiet, anvender vurderingskriterier ud fra, hvad man kunne kalde et nypositivistisk paradigme med vægt på intern gyldighed. Mens de organisationer, der ikke tager udgangspunkt i evidenshierarkiet, betoner primærstudiernes relevans for det specifikke emne og vurderer dem på deres eget designs præmisser så at sige. Med hensyn til syntetisering af resultaterne anbefaler organisationer, der prioriterer RCT (som er toppen i evidenshierarkiet) en metaanalyse som den ideelle metode. De andre organisationer har en pluralistisk tilgang til syntetisering og anbefaler også narrativ og konceptuel syntetisering afhængig af problemstilling, og hvilke design primærstudierne har anvendt). I nogle forskningsoversigter kombineres flere metoder til syntetisering.

For det tredje giver vi, ud fra eksempler på forskningsoversigter og andre kilder, en analyse af, hvordan organisationer, der udarbejder evidens faktisk praktiserer deres egne retningslinjer og anbefalinger, hvad angår metoder. Her viser det sig, at de organisationer, der lægger vægt på evidenshierarkiet, rent faktisk også har en højere andel forskningsoversigter, som alene medtager RCT-baserede primærstudier. Men selv de varmeste fortalere for RCT producerer også forskningsoversigter, hvor kvasi-eksperimentelle og andre design er medtaget. En forklaring herpå er, at der simpelthen ikke er tilstrækkelige RCT af god kvalitet på det aktuelle område. Det gælder især for Europa, hvor RCT-baseret forskning på social- og uddannelsesområdet er mindre udbredt end i USA.

Eftersom metodedebatten i og omkring evidensbevægelsen kredser om RCT som toppen af evidenshierarkiet, giver vi for det fjerde en oversigt over argumenter for og imod RCT som forskningsdesign. RCT er et design, der er velegnet til at analysere effekter af afgrænsede og specificerede interventioner, fx kliniske forsøg. Ved at benytte lodtrækning til indsats- og kontrolgruppen sikres det, at såvel dem, der modtager indsatsen, som dem, der leder undersøgelsen, ikke ved, hvem der indgår i henholdsvis indsats- og kontrolgruppe (blinding). Desuden kan alle faktorer holdes konstant ved at udarbejde baselinemålinger før indsatsen påbegyndes og ved at gennemføre eftermålinger af effekterne af indsatsen. Der er imidlertid også udfordringer og begrænsninger ved at anvende RCT. For det første producerer RCT-design smal evidens i den betydning, at de alene har udsagnskraft om effekter, dvs. om, hvilke interventioner der virker henholdsvis ikke virker. De har ingen udsagnskraft om, hvorfor noget virker eller ikke virker og ej heller om, hvordan modtagerne oplever indsatsen. For det andet er der i nogle sammenhænge en række tekniske problemer. Når RCT anvendes på velfærds- og uddannelsesområdet, er det fx ofte svært at sikre blinding, dvs at forsøgspersonerne ikke ved, om de er i kontrol- eller forsøgsgruppen. Herudover har kritikere formuleret en række argumenter mod at anvende RCT på områder, hvor interventioner er sammensatte og dynamiske, og hvor konteksten har en betydning for, om og hvordan interventionerne virker. Diskussionerne om styrkerne og svaghederne ved at anvende RCT viser variationer i kausalitetsforståelse og videnskabsteoretiske paradigmer.

Endelig giver vi, for det femte, en kort introduktion til begrebet »evidenstypologi« som alternativ eller supplement til evidenshierarkiet. Tankegangen bag udarbejdelse af evidenstypologier er, at forskellige undersøgelsesdesign har potentiale til at besvare forskellige typer af undersøgelsesspørgsmål. Frem for at tage afsæt i en forestilling om, at nogle undersøgelsesdesign er stærkere end andre, er udfordringen at tilpasse undersøgelsesdesign til den problemstilling, der ønskes belyst. Typologitænkningen kan inspirere til at udarbejde mere helhedsorienterede undersøgelsesdesign og forskningsoversigter, hvor viden om forskellige aspekter af givne indsatser vurderes med afsæt i en vifte af undersøgelsesdesign.

2 Indledning

2.1 Formål og afgrænsning

Dette er en rapport om evidensbevægelsens metoder. Formålet med rapporten er todelt. For det første at give en beskrivelse af det grundlag (vurderingskriterier), der anvendes ved vurdering af primære studier, når der foretages udvælgelse af studier, der skal indgå i systematiske forskningsoversigter. For det andet at fremlægge argumenter for og imod de forskellige vurderingskriterier, som det fremgår af nyere debat inden for evidensbevægelsen.

Vores udgangspunkt er de specialiserede, evidensproducerende organisationer, som er oprettet siden starten af 1990'erne. Vi afgrænser os hovedsageligt til sundheds-, social- og uddannelsesområdet, men inddrager også kilder fra det kriminologiske område. Geografisk afgrænses primært til evidensproducerende organisationer i Europa, men litteratur og eksempler fra andre regioner (især USA) medtages også.

2.2 Evidensbevægelsens baggrund

Begrebet evidens har i de senere år tiltrukket sig fornyet og stærkt voksende interesse. Evidensbaseret politik, evidensbaseret forvaltning, evidensbaseret ledelse, evidensbaseret praksis (medicin, velfærd, socialt arbejde, uddannelse mv.), ja evidensbaseret alt-muligt er blevet et plusord ikke blot i forskningskredse, men også i politik, forvaltning og professionel praksis.

Evidensbasering handler om at udarbejde politik, forvaltning, praksis etc. med afsæt i den bedst mulige viden om, hvilke indsatser der virker henholdsvis ikke virker. Evidensbevægelsen sætter således fokus på effekter af indsatser og interventioner og arbejder for at syntetisere allerede foreliggende viden samt synliggøre den aktuelt bedste viden og stille denne til rådighed for beslutningstagere i politik og praksis. Evidensbevægelsens rationale kan i forlængelse heraf siges at være forankret dels i et ønske om at bidrage til, at der ikke bruges ressourcer på interventioner, der ikke er belæg for virker, dels at synliggøre gode standarder for professionel praksis, som aktuel praksis kan vurderes op imod. Med beslutningsteoretiske termer kan man karakterisere evidensbevægelsens tænkning som båret både en fejlrettelsestænkning og en mere formålsrationel benchmarkingtænkning.

Evidensbevægelsens idemæssige grundlag kan føres tilbage det 19. århundrede (Sackett 1997) og har endda rødder helt tilbage til tanker om udførelsen af kontrollerede forsøg inden for det medicinske område i 1700-tallet (Chalmers 2001). Et vigtigt fundament for bevægelsen etableres i 1930'erne og 1940'erne, da man eksperimenterer med kontrollerede lodtrækningsforsøg inden for det medicinske område (Oakley 2000). Bevægelsen får dog først for alvor fat i løbet af 1970'erne, hvor epidemiologen Archie Cochrane i en banebrydende publikation argumenterer for nødvendigheden af at prioritere de begrænsede ressourcer inden for sundhedsvæsenet på baggrund af fair test (Cochrane 1999|1972). Med andre ord: Sundhedsfaglig prioritering skal være evidensbaseret. Parallelle bestræbelser foregår inden for socialområdet, kriminologi og uddannelsesområdet med udgangspunkt i amerikaneren Donald Campbells tænkning (Campbell 1969).

Et centralt virkemiddel i evidensbevægelsen er de såkaldte systematiske forskningsoversigter, også kaldet systematiske reviews. Et systematisk review søger at identificere alle relevante studier (primærstudier), der omhandler en given problemstilling. Efter en kritisk vurdering af primærstudiernes kvalitet besluttes det at inkludere nogle i reviewet, mens andre ekskluderes. Resultaterne af de inkluderede primærstudier syntetiseres. Dette kan fx ske statistisk (via metaanalyse) eller narrativt (kaldes undertiden en narrativ forskningsoversigt, Petticrew og Roberts 2006). Systematiske re-

views betragtes som særligt valide, fordi de sammenfatter resultaterne af større datamaterialer. Endvidere betragtes de som et middel til at skabe overblik i den hastigt voksende jungle af forsknings- og evalueringsresultater.

Evidensbevægelsen er i dag organisatorisk forankret i en lang række internationale såvel som nationale organisationer. Den nok mest produktive er Cochrane Collaboration på det medicinske område, der siden sin oprettelse i 1993 har produceret omkring 4000 systematiske reviews (inklusive igangværende projekter). Organisationen har desuden dannet skole for en lang række organisationer inden for det medicinske, det sociale og det kriminologiske område såvel som uddannelsesområdet. Ligeledes på det medicinske område findes nationale organisationer for medicinsk teknologivurdering. Disse organisationer søger typisk på et evidensbaseret grundlag at vurdere en bred vifte af aspekter ved nye og eksisterende teknologier til anvendelse i sundhedsvæsenet.

De nordiske lande, især Sverige og Danmark, har spillet en aktiv rolle i evidensbevægelsens organisatoriske udvikling. Således fik man allerede i 1993 et regionalt Cochrane-center ved Rigshospitalet i København, ligesom der i 2001 blev oprettet et center under den tilsvarende organisation på socialområdet, Campbell Collaboration, ved Socialforskningsinstituttet. Endvidere har man siden 1997 haft et institut for medicinsk teknologivurdering under Sundhedsstyrelsen, CEMTV[1], mens der under Lægemiddelstyrelsen siden 1999 har eksisteret et Institut for Rationel Farmakoterapi, som på et evidensbaseret grundlag søger at sikre den mest rationelle udnyttelse af lægemidler i Danmark. I Sverige har der eksisteret evidensbaserede organisationer på det medicinske område siden 1987, hvor man oprettede Statens Beredning för Medicinske utvärderingar, og på det sociale område siden 1993, da Centrum för utvärdering av socialt arbejde i dag Instituttet för utveckling av metoder i socialt arbete blev en realitet. I Norge og Finland blev evidensbevægelsen organisatorisk rodfæstet i midten af 1990'erne på det medicinske område og i slutningen af samme årti på socialområdet. Evidensbevægelsens historie og organisatoriske forankring er nærmere beskrevet i Bhatti, Hansen og Rieper (2006).

Evidensbevægelsen er stadig hastigt voksende. Internationalt øges intensiteten i vidensakkumuleringen i de evidensproducerende organisatio-

ner, ligesom der fx i OECD-regi pågår et arbejde, der søger at diskutere og yderligere udbrede fænomenet inden for uddannelsesområdet. Med OECD i rollen som idespreder blev der i Danmark i 2006 etableret et clearinghouse – et evidensproducerende center for uddannelse – på Danmarks Pædagogiske Universitetsskole på Aarhus Universitet. Det er således vigtigere end nogensinde at forstå og derved kunne diskutere evidensbevægelsen og dens grundlag.

2.3 Problemstilling: evidenshierarkiet og evidensdebatten

Som nævnt er udarbejdelsen af systematiske reviews – dvs. forskningsoversigter der syntetiserer viden fra multiple primærstudier – et helt centralt træk ved evidensbevægelsen. Som nævnt indebærer udarbejdelsen af reviewene en kritisk litteraturvurdering, hvor man inkluderer nogle primærstudier i reviewet, mens andre ekskluderes. Vi vil i denne rapport se særligt på, hvilke kriterier der ligger til grund for denne udvælgelse. Inklusionsspørgsmålet er overordentlig interessant, eftersom det reflekterer, hvad der i evidensbevægelsen bliver betragtet som valid og legitim viden, og hvilke vidensformer der omvendt bliver betragtet som mindre valide. Dette er om noget et centralt spørgsmål i vidensamfundets hastige vidensproduktion og -akkumulering.

Udgangspunktet for udvælgelsen af primærstudier til systematiske reviews er opfattelsen af, at ikke alle vidensformer er lige gode og således må rangordnes. Rangordningen sker ofte ved hjælp af det såkaldte evidenshierarki, som angiver, hvilke former for viden der er at foretrække frem for andre. Som den generelle evidensdebat er metodedebatten om evidenshierarkiet særdeles intensiv og ofte ganske polariseret. Den er endvidere til tider ganske teknisk og kan derfor forekomme svært tilgængelig. Debatten er imidlertid helt central, da evidensbevægelsen ofte legitimerer sig med udgangspunkt i videnskabelig stringens og validitet, hvilket netop er det, der diskuteres i debatten. Der er således et betydeligt behov for at præsentere debattens vigtigste argumenter og søge at give en vurdering af deres fundament. Dette vil være hovedformålet med nærværende rapport.

Rapporten vil mere præcist søge at besvare følgende spørgsmål, der omhandler evidenshierarkiet i sig selv, dets videnskabsteoretiske grundlag såvel som nogle af de praktiske problemstillinger, der kan knytte sig til det, når det anvendes i praksis:

1. Hvad er evidenshierarkiet?
2. Hvilke argumenter knytter der sig typisk for og imod det?
3. Hvordan træffes konkrete beslutninger om inklusion og eksklusion – hvor i hierarkiet skal snittet lægges?
4. Hvordan kvalitetsvurderes primærstudier, og hvordan syntetiseres resultaterne herfra?
5. Hvilket videnskabsteoretisk grundlag knytter sig til argumenterne i evidensdebatten?

2.4 Anvendte metoder og kilder

Rapporten bygger på en række forskellige kilder.[2] For det første har vi gennemgået evidensorganisationernes vejledninger til kvalitetsvurdering af primærstudier. Vi har med andre ord set på, hvordan organisationerne på det generelle niveau præsenterer deres reviewpraksis. Herudover har vi i relation til de vigtigste organisationer analyseret deres reviewpraksis på handlingsplanet. Vi har gennemgået (et udvalg af) systematiske reviews og registreret, hvordan praksis konkret er foregået. Dette dokumentariske materiale er suppleret med observation på konferencer og interview med nøglepersoner. Endelig trækker vi på den hastigt voksende internationale faglitteratur om evidensbevægelsen og den metodologiske litteratur om reviewpraksis.

Rapporten er opdelt i seks afsnit. I afsnit 3 præsenteres tankegangen bag evidenshierarkiet. De forskellige undersøgelsesdesigns omtales kort, og der gives eksempler på evalueringer, der har benyttet disse. I afsnit 4 præsenteres de håndbøger og guidelines for udarbejdelse af systematiske reviews, som er udarbejdet af de væsentligste evidensproducerende organisationer. Heraf fremgår det også, hvilken politik organisationerne har for hvilke typer af designs**,** der bør in- henholdsvis ekskluderes fra reviews. I afsnit 5 ser vi herefter på, hvordan organisationerne rent faktisk praktiserer reviewarbejdet, er der med andre ord overensstemmelse mellem deres poli-

tik for og tale om reviewarbejdet og den måde, der handles på. Afsnit 6 indeholder diskussionen af argumenter for og imod anvendelse af RCT som guldstandard, og afsnit 7 præsenterer og diskuterer alternativer til evidenshierarkiet de såkaldte evidenstypologier.

3 Evidenshierarkiet: En rangorden af forskningsdesign

Evidensbevægelsens udgangspunkt er som allerede nævnt, at ikke alle vidensformer er lige valide. Evidenshierarkiet repræsenterer således en rangordning af den viden, som resultaterne fra primærstudier udgør, og som ligger til grund for systematiske reviews og i sidste ende deres anbefalinger for politisk såvel som praktisk handlen. Rangordningen tager udgangspunkt i det *forskningsdesign,* som er anvendt i primærstudierne. Kvalitetsvurderingen af de enkelte primærstudier foregår således dels ved at identificere, hvilket design der er anvendt, og fastlægge om det pågældende design ligger over eller under et fastlagt snit (cut) i rangorden af design, dels ved at gennemføre en kvalitetsvurdering af designet i de primærstudier, der ligger over det fastlagte snit. Rangordenen synes foretaget efter det grundprincip, der bygger på, at kausalitet (årsag-virkningsrelationer) bør afdækkes ved at udelukke den kontrafaktiske problemstilling: hvis andre mulige årsager til virkningen kan udelukkes, så er den undersøgte årsag (indsatsen) gældende. Udelukkelsen af andre mulige årsager (og isoleringen af den undersøgte årsag) foretages ifølge kausalitetsopfattelsen i den empirisk-analytiske videnskabsteoretiske tradition bedst ved, at der anvendes et randomiseret kontrolleret eksperiment-design (RCT). Andre designs er svagere til at håndtere det kontrafaktiske problem, men fx statistiske multivariate analyser, hvor den undersøgte årsag isoleres ved at »holde andre mulige årsagsfaktorer konstante«, afspejler RCT-tankegangen blot i en statistisk analyse. Ud fra denne tradition er andre design meget svagere. Fx tilkendes brugernes observationer og vurderinger af, hvordan en indsats virker på dem, ringe

vægt, idet brugerne antages at have en forvredet eller mindre gyldig opfattelse af kausalitet.

Der findes ingen entydig etableret konsensus omkring den eksakte rangordning og antallet af forskningsdesign. For eksempel opdeler nogle forfattere hierarkiet i kun 2-3 kategorier af forskningsdesigns, mens andre anvender 8-9. Ligeledes er der betydelig kontrovers omkring den relative placering af ikke-randomiserede eksperimentelle design og kohortestudier. Et tredje stridspunkt er, hvor finopdelt bunden af evidenshierarkiet skal være. Nogle samler ekspertvurderinger, eksempler på god praksis (til tider benævnt anekdotisk viden) mv. i én kategori, mens andre også på dette niveau foretager en nøjere rangering.

Et fællestræk ved alle udgaver af evidenshierarkiet er dog, at de giver fortrinsstilling til det randomiserede kontrollerede forsøg (oftest benævnt ved forkortelsen RCT – *randomized controlled trial*), på dansk også betegnet lodtrækningsforsøg. Den mest udbredte »klasse« af rangordninger er inspireret af epidemiologen David Sackett, der regnes som en af evidensbevægelsens absolutte foregangsmænd, og ser med små variationer ud, som det fremgår af figur 3.1.

Figur 3.1 Det fulde evidenshierarki

Niveau	Studietype
1a 1b 1c	Systematiske reviews af RCT. Enkeltstående RCT af god kvalitet. Kontrollerede, men ikke randomiserede forsøg.
2a 2b	Systematiske reviews af kohortestudier. Enkeltstående kohortestudier. Dårlige RCT.
3a 3b	Systematiske reviews over casekontrolstudier. Enkeltstående casekontrolstudier.
4	Caseserier eller kohortestudier eller casekontrol af dårlig kvalitet.
5	Ekspertvurderinger, konsensuskonferencer, kvalitative designs mv.

Kilder: Frit efter Sackett et al. (2000); Pedersen et al. (2001).

Som det fremgår af figur 3.1, består evidenshierarkiet af fem overordnede kategorier, der er opdelt efter forskningsdesign. Inden for hver af de tre forskningsdesign, der betragtes som de bedste, er der en underopdeling, således at der skelnes mellem systematiske reviews og enkeltstående studier med et bestemt forskningsdesign. Systematiske reviews vægtes højere, da datagrundlaget typisk er større og af mindst samme kvalitet som for

de enkeltstående undersøgelser med samme forskningsdesign. Dette er også årsagen til, at de ledende organisationer i evidensbevægelsens søger at forestå udarbejdelsen af sådanne reviews. Da vi imidlertid i denne rapport primært ønsker at se på inklusionen af primærstudier til systematiske reviews, »renser« vi evidenshierarkiet og præsenterer i figur 3.2 alene rangordenen for design af primærstudier.

Figur 3.2 Evidenshierarki: To eksempler på engelsk og en udgave på dansk

Niveau	Dansk udgave	Eksempel A Hierarchy of evidence in meta-analysis (Pawson 2006, p49)	Eksempel B Hierarchy of evidence (Clarke 2006, p562f, based on Stevens & Abrams 2001)
1	Randomiserede, kontrollerede eksperimenter (dobbeltblindede, enkeltblindede eller ublindede)	Randomized controlled trials (with concealed allocation)	At least one properly designed RCT of appropriate size
2	Kvasieksperimenter: Kontrollerede forsøg baseret på matching	Quasi-experimental studies (using matching)	Well-controlled trials without randomisation
3	Forløbsstudier	Before-and-after comparison	Well-designed cohort or case control studies
4	Tværsnitsundersøgelser	Cross-sectional, random sample studies	Multiple time series or dramatic results from uncontrolled experiments
5	Procesevaluering, aktionsforskning o.lign.	Process evaluation, formative studies and action research	Opinions of respected authorities based on clinical evidence, descriptive studies or expert committee
6	Kvalitative casestudier og etnografiske feltstudier	Qualitative case studies and ethnographic research	Small uncontrolled case series and samples
7	Erfaringer og eksempler på god praksis	Descriptive guides and examples of good practice	
8	Ekspertvurderinger	Professional and expert opinion	
9	Brugervurderinger	User opinion	

Figur 3.2 præsenterer to eksempler på evidenshierarkier. Eksempel A, der også præsenteres i dansk oversættelse, indeholder 9 niveauer, eksempel B 6 niveauer. Bemærk, at i eksempel B på niveau 1, 2 og 3 vurderer man også, hvor godt de enkelte design er gennemført. Det ses endvidere, at eksempel A og B adskiller sig ved ikke at indeholde sammenfaldne begre-

ber for design. Herudover rangordnes eksperters meninger i A efter kvalitative casestudier og etnografiske studier, mens de i B rangordnes før mindre, ikke kontrollerede casestudier.

Da diskussionen inden for evidensbevægelsen hovedsageligt har handlet om, hvor snittet i evidenshierarkiet skal ligge, altså hvilke forskningsdesign der vurderes som acceptable, vil nedenstående gennemgang af design have hovedvægt på de øverste niveauer i evidenshierarkiet, hvor der også gives eksempler på de enkelte design, mens de nederste niveauer behandles mere kortfattet.

3.1 Det randomiserede kontrollerede forsøg

Det randomiserede kontrollerede forsøg (RCT'et) betragtes ifølge evidenshierarkiet som det mest valide forsknings- og evalueringsdesign og anses derfor for at være evidensbevægelsens guldstandard (Sackett et al. 1996; Concato et al. 2000). Et RCT indebærer opdeling af undersøgelsessubjekter i mindst en interventionsgruppe og en kontrolgruppe. Subjekterne (som ofte forstås som individer, men som også kan være organisationer eller andre enheder) i interventionsgruppen modtager den indsats, man ønsker at undersøge, mens kontrolgruppens subjekter modtager sammenligningsgrundlaget, som fx kan være placebo eller den hidtidigt tilbudte indsats. Centralt i forskningsdesignet er, at allokeringen til disse to grupper er randomiseret (tilfældig), således at man eliminerer selektionsbias og sikrer, at indsats- og kontrolgruppen har de samme karakteristika (inden for statistisk tilfældige grænser). Med andre ord søger man ved randomiseringen at sikre, at kun ét karakteristikum er forskelligt i indsats- og kontrolgruppen, nemlig den intervention, man ønsker at undersøge (Madsen og Andersen 2005). Kontrolgruppen skal altså filtrere alle forudsete og uforudsete faktorer ud, når man til slut i forsøget sammenligner interventionseffekterne. Kontrollen er således stærkere end den, man kender inden for samfundsvidenskabens most-similar strategi, hvor der alene kontrolleres for kendte faktorer (Peters 1998). Sagt på en lidt anden måde: Guldstandarden indebærer, at indsats- og kontrolgruppen består af to ens grupper, der behandles samtidig, det såkaldte parallelle design. En undersøgelse, der er designet som et RCT, kan afdække, om

en indsats har ført til resultater, men den kan ikke i sig selv belyse, hvorfor eller hvorfor ikke indsatsen giver resultater.

Et vigtigt forhold i RCT-forskningsdesignet er graden af blinding, som angiver, hvorvidt forsøgssubjekterne og forsøgslederne er klar over, hvilke patienter/klienter der befinder sig i henholdsvis interventions- og kontrolgruppen. Hvis alle involverede parter har kendskab til allokeringen, tales der om ublindede forsøg. Hvis kun én af parterne, dvs. enten deltagerne eller forsøgslederne, kender tilskrivningen, kaldes det et enkeltblindet forsøg. Endelig kalder man de tilfælde, hvor allokeringen er skjult for begge parter, for dobbeltblindede (Sackett et al. 1997). Dobbeltblindede forsøg er normalt at foretrække, idet de anses som en beskyttende faktor mod menneskeskabt bias (Gøtzsche 1990), idet indsatsen rettet mod patienterne/klienterne såvel som evalueringen af outcome sker uden kendskab til den modtagne indsats (Day og Altman 2000). Ud over begreberne enkeltblinded og dobbeltblinded anvendes til tider begrebet tredobbeltblinded. Dette begreb refererer til forsøgsdesign, hvor det ikke er muligt at identificere om individuelle patienter/klienter har tilhørt interventions- eller kontrolgruppen, før efter de indsamlede data er blevet analyseret. Der findes en betydelig debat omkring det etisk forsvarlige i blinding, som vi vil vende tilbage til i forbindelse med diskussionen i afsnit 6 af de vigtigste argumenter for og imod evidenshierarkiet.

En særlig form for RCT er de såkaldt cross-over-studier, hvor forsøgspersonerne udgør deres egen kontrol ved at skifte gruppe undervejs. Cross-over-studierne kan ligesom de almindelige RCT være blindede, men er det ikke nødvendigvis. Studietypen accepteres nogle gange på lige fod med almindelige RCT (Madsen og Andersen 2005).

Det randomiserede kontrollerede forsøgs styrke er, at randomiseringen i en interventions- og en kontrolgruppe tager højde for tilstedeværelsen af ukendte faktorer, som kan have indflydelse på forsøgsresultatet (konfoundere). Man kontrollerer derved »automatisk« for alle faktorer og ikke alene for de faktorer, som er forskerne bekendt, hvorfor RCT betragtes som højere rangerende end fx økonometriske metoder (Sherman 2003). Den hyppige anvendelse af blinding beskytter endvidere som nævnt mod menneskelige bias. Anvendelsen af RCT-design sikrer den bedst mulige dokumentation af årsags-virknings-forholdet mellem indsats og effekt. Den

interne validitet karakteriseres på denne baggrund som værende meget høj. Netop disse forhold er også årsagen til, at evidensbevægelsen ofte betragter RCT som guldstandarden.

Kritikere har heroverfor fremført, at den eksterne validitet til gengæld ofte er lav, at der med andre ord kan være problemer med at generalisere resultaterne til andre situationer og kontekster (Launsø og Gannik 2000). Der er dog betydelig uenighed omkring dette (Chalmers 2005). Også dette vil blive uddybet i afsnit 6.

Et eksempel på et randomiseret kontrolleret forsøg på det medicinske område

Formålet var at undersøge effekten af akupunktur sammenlignet med minimal akupunktur og med ingen akupunktur hos patienter med spændingshovedpine. Der blev anvendt et treleddet randomiseret kontrolleret design. Forsøget foregik på klinikker i Tyskland og omfattede 270 patienter med periodevise eller kronisk spændingshovedpine. Interventionen bestod i 12 sessioner per patient over 8 uger og blev foretaget af en speciallæge.

Outcome målet var antal dage med hovedpine mellem hhv. 4 uger før randomisering og uge 9-12 efter randomisering, som patienterne havde beskrevet det i deres dagbøger. Resultatet viste, at antal af dage med hovedpine faldt med 7,2 dage i akupunkturgruppen sammenlignet med 6,6 dage i gruppen, der fik minimal akupunktur, og 1,5 dage i gruppen, der var på venteliste. Forfatterne konkluderer, at den akupunktur, der blev anvendt i dette forsøg, var mere effektivt end ingen behandling, men ikke signifikant mere effektivt end minimal akupunktur for behandling for spændingshovedpine (BMJ 2005;331:376-382 (13 August), doi:10.1136/bmj.38512.405440.8F (published 29 July 2005; Trial registration number ISRCTN9737659).

Et eksempel på et randomiseret kontrolleret forsøg på det sociale område

Nogle kommuner i Danmark er begyndt at tilbyde PMT (parent management training) behandling til familier med børn med adfærdsproblemer. Center for Anvendt Sundhedstjenesteforskning og Teknologivurdering

(CAST) på Syddansk Universitet gennemfører en evaluering af effekten af PMT. Evalueringen er tilrettelagt som en præ-post undersøgelse designet som et randomiseret kontrolleret forsøg. Blandt familier, der har erklæret sig villige til at indgå i forsøget, trækkes der lod om, hvorvidt de skal indgå i interventionsgruppen, der tilbydes PMT, eller i kontrolgruppen, der tilbydes den behandlingsform, deres kommune hidtil har benyttet. I begge grupper gennemføres før- og eftermåling, førmåling inden behandlingen iværksættes og eftermåling 10 måneder senere. Resultaterne af undersøgelsen forventes at foreligge ultimo 2008. Eksemplet illustrerer, at det på det samfundsvidenskabelige område kan være umuligt at designe RCT som blinde forsøg. Både behandlere og familierne ved, hvilken gruppe den enkelte familie indgår i, når først lodtrækningen har fundet sted.

3.2 Det ikke-randomiserede, kontrollerede forsøg: Matching

Også kvasieksperimenter defineret som kontrollerede forsøg baseret på matching er tilrettelagt med en forsøgs- og en kontrolgruppe. Allokeringen til grupperne er imidlertid ikke randomiseret via lodtrækning, men grupperne er sammensat, så de matcher hinanden, dvs. er ens på de variable, der formodes at være centrale. Forskningsspørgsmålet afgør, hvilke variable der matches på. Anvendelse af kvasieksperimenter baseret på matching forudsætter derfor en teori om, hvilke årsager der ligger bag forandringer i det felt, der analyseres.

Kvasieksperimenter betragtes sædvanligvis som værende lavere i evidenshierarkiet end RCT. Dette skyldes, at den manglende randomisering kan indføre bias i allokeringen af forsøgssubjekter (selektionsbias), således at interventionsgruppe og kontrolgruppe reelt ikke bliver sammenlignelige.

Den eksisterende litteratur om bias på det medicinske område tyder på, at selektionsbias i praksis næsten altid vil favorisere den undersøgte intervention (Schulz et al. 1995; Gluud 2005) og derved potentielt legitimere indsatser, der ikke har effekt. Dette er i evidensbevægelsens optik problematisk, da en central motivation for at udføre fair test er, at selv velmenende interventioner kan gøre mere skade i gavn (Chalmers 2003).

De randomiserede kontrollerede forsøg og de ikke-randomiserede kontrollerede forsøg udgør tilsammen en hovedgruppe af design, der kan kaldes kontrollerede eksperimentelle studier, idet det er forsøgslederne, der styrer, hvilken eksponering forsøgssubjekterne er udsat for (Coggon et al. 1997).

Naturlige eksperimenter er en anden kategori af ikke-randomiserede eksperimenter. De kaldes naturlige, fordi indsats- og kontrolgruppe ikke er frembragt som led i undersøgelsen, men er opstået som følge af andre forhold, fx ved lovgivning, administrative eller naturlige forhold. Et eksempel på et naturligt experiment, som analyseres ved økonometriske modeller og metoder, er en undersøgelse af aktiveringsindsatser over for ledige. Princippet er kort sagt at udnytte, at de forskellige aktiveringsindsatser kan finde sted på forskellige tidspunkter i kontanthjælpsperioden. Man modellerer aktiveringsindsatserne som forklarende variable, der varierer over tid. Man kan hermed undersøge både fastlåsningseffekter og aktiveringseffekter af indsatserne. Fraværet af kontrolgruppe indebærer, at der ikke umiddelbart kan måles absolutte effekter af en indsats, men derimod relative effekter af forskellige indsatser (Rosholm 2004).

Et eksempel på et ikke-randomiseret kontrolleret forsøg på det medicinske område

Et studie sammenlignede patienter med overfølsomhedssygdomme, som havde fået hhv. etableret behandling hos praktiserende læger og behandlinger hos klassiske homeopater. Patienterne valgte selv behandling. Patienternes sundhedstilstand var ensartet ved starten af behandlingerne. Behandlingsresultaterne blev målt ved patienternes retrospektive, selvrapporterede vurdering af effekter. Resultatet var, at begge grupper oplevede en forbedret psykologisk tilstand efter behandling, men dobbelt så mange patienter i klassisk homeopatisk behandling oplevede forbedringer i deres sundhedstilstand i forhold til patienter i etableret behandling. Logistisk regressionsanalyse, hvor man bl.a. tog hensyn til patienternes erhverv, viste, at behandlingerne var den eneste signifikante uafhængige variabel (Launsø m.fl. 2006).

Et eksempel på et ikke-randomiseret kontrolleret forsøg på det samfundsvidenskabelige område: Kampagne for at anvende cykelhjelme

I Frederikssund blev der i en periode på godt to år gennemført en kampagne, som skulle begrænse antallet af ulykker. Alle typer ulykker og alle aldersgrupper var målgruppen. Der var ønske om at måle effekten af kampagnen med hensyn til ændringer af viden, holdning og adfærd. Dette skulle være et supplement til en detaljeret registrering af skader behandlet på skadestuen.

Det blev valgt bl.a. at måle effekten på børn i 6. klasse. Spørgeskemaer blev udfyldt før og efter kampagnen. Det samme skete i Køge – som således fungerede som en sammenligningsgruppe, som kunne afsløre, om der var nogle generelle tendenser for hele landet, fx i retning af øget brug af cykelhjelme. Undersøgelserne dækkede således samme alderstrin i før- og efterrunderne, dvs. at det drejede sig om tværsnitsundersøgelser, hvor alle på det pågældende trin i de pågældende kommuner blev medtaget. I alt blev 1.045 børn spurgt. Kampagnens effekt viste sig at være begrænset – der var få spørgsmål, som udviste en markant ændring. I visse tilfælde fandtes også ændringer, som gik i negativ retning – set med forebyggelsesøjne. Fx var der flere, der kørte med cykelhjelm, men der var også markant flere, som syntes, at det var sjovt at køre hurtigt på cykel.

Det viste sig endvidere svært at anvende sammenligningsgruppen til sit egentlige formål, nemlig at vise, om de ændringer, som skete i Frederikssund, skyldtes generelle landsdækkende forhold. Problemet er, at de to kommuner er ganske forskellige, når man kommer ned på enkelte spørgsmål, som fx hvordan børnene vurderer skolepatrulje og gårdvagt. Når udgangspunktet er forskelligt, er det svært at tolke ændringerne i sammenligningsgruppen.

Den største nytte af sammenligningsgruppen var det store datamateriale, som blev udnyttet til at beskrive sammenhæng mellem baggrundsvariabler og variabler, som ikke var påvirket af kampagnen. Til dette kunne hele datamaterialet benyttes. Mehlby, Rieper og Togeby 1993, ss. 59-60).

3.3 Forløbsundersøgelser

Ud over RCT-design og kvasieksperimenter baseret på matching findes der en vifte af andre typer af undersøgelsesdesign. Da der benyttes varierende terminologi på fx det medicinske og det samfundsvidenskabelige område, kan det være lidt svært at skabe det forkromede overblik. Vi forsøger nedenfor at give vores bud.

Det tredje niveau i evidenshierarkiet udgøres af forløbsstudier (longitudinal design), på dansk i nogle sammenhænge også benævnt længdesnitsundersøgelser eller blot længdeundersøgelser. I forløbsundersøgelser følges de samme individer eller organisationer over tid. Der findes flere typer af forløbsstudier: Kohortestudier, casekontroldesign og før-efter-sammenligninger.

I kohorteundersøgelser observeres undersøgelsesobjekternes tilstand over tid, men de udsættes ikke for interventioner, der er indført som en del af undersøgelsen (Andersen og Osler 2004). En kohorte forstået som en gruppe mennesker, der deler et fælles karakteristika, fx er født i samme uge, eller har været ude for en fælles hændelse, fx et hjerteanfald, følges over tid med sigte på at afdække deres udvikling.

Større folkeundersøgelser tilrettelægges ofte som kohortestudier. Sådanne kan tage udgangspunkt i registerdata eventuelt kombineret med spørgeskemaundersøgelser. I kohorteundersøgelser indgår oftest langt flere individer end i forbindelse med eksperimentelle studier. Dette skylder dels, at de typisk er billigere at gennemføre pr. forsøgsperson, da man alene observerer og ikke intervenerer, dels at der er behov for langt større datasæt for at generere meningsfulde resultater, da man ikke som i RCT kan isolere en enkelt variabel. Danmark har relativt omfattende registre og har således gode forudsætninger for at lave kohortestudier. Inden for kohortestudier skelnes normalt mellem prospektive og retrospektive design, hvor man i førstnævnte starter med at undersøge en populationen for at følge dens udvikling på en række variabler, mens man i sidstnævnte alene undersøger den efter en given hændelse/eksponering (Andersen og Sørensen 1997). I retrospektive design sammensætter og sammenligninger man ofte to grupper jf. eksemplet vedrørende hjerteanfald nedenfor.

I casekontrolstudier ønsker man som i kohortestudier at undersøge årsager til tilstande i undersøgelsessubjekternes eksponering til forskellige

faktorer. I modsætning til kohortestudierne udvælger man dog ikke undersøgelsessubjekterne med udgangspunkt i risikofaktorerne, men derimod om patienterne har den tilstand, man ønsker at undersøge (casegruppen) eller ikke (kontrolgruppen). Årsagerne til det forskellige outcome mellem de to grupper undersøges derefter statistisk med baggrund i oplysninger om deres forskellige eksponering (Coggon et al. 1997). Casekontrol studier er typisk associeret med problemer omkring identificering af en passende kontrolgruppe, ligesom designet i endnu mindre grad end kohortestudierne eliminerer uforudsete sammenhænge (Coggon et al. 1997).

I før-efter-sammenligninger, til tider benævnt refleksiv kontrol (Vedung, 1998: 140), undersøges en interventionsgruppe før og efter en intervention. Formålet er at vurdere, om der er sket forandringer på de dimensioner, interventionen sigter mod at forandre. Før-efter-sammenligninger kan tilrettelægges enten som serieoplæg, dvs. med flere målinger før og efter interventionen, eller mere simpelt med en enkelt måling før og efter. Før-efter-sammenligninger kan anvendes, hvis den indsats, der analyseres, er iværksat for alle, hvis det med andre ord ikke er muligt at arbejde med en kontrolgruppe. Den principielle svaghed ved dette design er, at selv om det viser sig, at der er sket forandringer, er der ingen sikkerhed for, at det er interventionen, der er årsagen. Kan der modsat ikke dokumenteres forandringer, er der ej heller sikkerhed for, at interventionen ikke har skabt forandringer. Problemet er, at andre forandringer end interventionen kan have påvirket situationen.

På det medicinske område benyttes også betegnelsen caseserier, som er et studie af en lille gruppe patienter, der udsættes for en påvirkning. Caseserierne er typisk karakteriseret ved at indeholde meget detaljeret information om den enkelte patients eksponering, kliniske tilstand og øvrige karakteristika i modsætning til de hidtidige nævnte design, der primært ser på større populationer og gennemsnitlige værdier af standardiserede outcomemål. I caseseriedesign anvendes der ikke kontrolgrupper. Der er nogen diskussion om caseseriers værdi, men de er generelt placeret lavt i evidenshierarkiet, fordi de ofte bliver betegnet som deskriptive og i bedste fald en primitiv form for casekontrolstudier (Cummings og Weiss 1998). Caseserier finder dog især anvendelse i forbindelse med sygdomme, der

forekommer sjældent, og i forhold til hvilke det derfor er umuligt at anvende en af de øvrige undersøgelsesformer.

På det samfundsvidenskabelig område benyttes også betegnelsen panelundersøgelser (Bryman 2004). I panelundersøgelser udvælges et panel bestående af personer, husholdninger, organisationer etc., og data indsamles fra panelet ved mindst to lejligheder. Panelet kan bestå af en tilfældigt udtræk af en større population.

Et eksempel på et forløbsstudie på det medicinske område

Et studie viser, at årsager til hjertesygdom er de samme over hele kloden. 15.000 mennesker i 52 lande, som havde haft deres første hjerteanfald, blev matchet med tilsvarende gruppe af mennesker af samme alder, køn og geografisk sted, som ikke havde haft hjerteanfald. Forløbet foregik over 10 år. Studiet har fået meget rosende bemærkninger, selv om det ikke er et RCT:

»This study confirms that the risk factors are the same all over the planet and... has made it possible to assess the weight of the different risk factors,« said Dr. Jean-Pierre Bassand, president of the European Society of Cardiology. It's a fantastic study. Editor of The Lancet medical journal where the findings have just been published: »Probably the most robust study on heart disease risk factors ever conducted.« Associated Press report, reporter Emma Ross, Aug 30, 2004.

Et eksempel på et forløbsstudie på det socialmedicinske område

Sundheds- og sygelighedsundersøgelser (SUSY):

Danskernes livsstil, sundhedsvaner og belastende levevilkår har en væsentlig del af ansvaret for helbredsproblemer og for tidlig død. Viden herom er af central betydning for at tilrettelægge en effektiv, forebyggende indsats. Statens Institut for Folkesundhed har det nationale ansvar for at gennemføre undersøgelser af danskernes sundhed og sygelighed. Der indsamles data om befolkningens sundhed og sygdom og om forhold af betydning herfor – fx risikofaktorer i livsstil og levekår. Undersøgelserne er baseret på repræsentative interview- og spørgeskemaundersøgelser i den voksne befolkning. Der indsamles såvel tværsnitsdata som kohortedata. Efterfølgende kobles data med en række registre.

3.4 Tværsnitsundersøgelser

I en tværsnitsundersøgelse (cross-sectional designs) indsamles data vedrørende mere end et case (normalt vedrørende mange cases) på et givent tidspunkt i tid. Der indsamles kvantitative eller kvantificerbare data vedrørende to eller flere variable (normalt flere end to). Sigtet er at belyse variation. Tværsnitsundersøgelser kan afdække samvariation mellem variable, men ikke kausalsammenhængenes retning, hvorfor deres resultater betragtes som havende lavere intern validitet end resultaterne af eksperimentelle design.

Et eksempel på en tværsnitsundersøgelse på det medicinske område
I projektet »Den gode medicinske afdeling«, der nu er integreret i det, der benævnes »Den Danske Kvalitetsmodel«, er der flere gange gennemført tværsnitsundersøgelser af de medicinske afdelingers kvalitet. I undersøgelserne sammenlignes afdelingerne på en række standarder knyttet til patientforløb, herunder planlægning af patientforløb, medicinering, udskrivelse, ernæring, genoptræning mv. Den enkelte tværsnitsundersøgelse giver afdelingerne mulighed for at sammenligne sig med andre afdelinger (benchmarking). Gentagne tværsnitsundersøgelser giver mulighed for at vurdere, om der sker en kvalitetsudvikling over tid.

Et eksempel på en tværsnitsundersøgelse på det sociale område
1. juli 2004 trådte en ny lov i kraft, der betød, at aldersgrænsen for salg af alkohol i detailhandelen blev hævet fra 15 til 16 år. Center for Alkoholforskning ved Statens Institut for Folkesundhed blev bedt om at evaluere konsekvenserne af den nye lovgivning. Evalueringen blev gennemført som to tværsnitsundersøgelser gennemført henholdsvis i maj-juni 2004 og maj-juni 2005. I begge runder blev 8000 tilfældigt udvalgte danskere i alderen 13 til 16 år udtrukket af CPR-registeret og tilsendt et spørgeskema om brug og køb af alkohol. Evalueringen viste, at de 15-åriges køb af alkohol var nedbragt med den nye lov, men samtidig at forbruget af alkohol blandt de 13-16 årige stort set var uændret (Jørgensen m.fl. 2006).

3.5 Procesevaluering, aktionsforskning o.l.

Procesevaluering er placeret relativt langt nede i evidenshierarkiet. Procesevaluering anlægger et historisk helhedsorienteret perspektiv. Evalueringsdesignet belyser indsatsens historiske baggrund, karakteristika ved indsatsen fx dens tekniske kompleksitet, implementeringsprocessen, herunder behandlernes ressourcer og holdninger, brugernes modtagelse af indsatsen, deres forståelse, holdninger og adfærd vis-a-vis behandlerne samt kontekstens betydning, herunder kontrollen af indsatsen, relationen til beslægtede indsatser, mediernes behandling af indsatsen mv. Procesevaluering har et potentiale til at forklare, hvordan indsatser udvikles over tid, samt hvordan denne udvikling og den kontekst, indsatsen leveres i, påvirker, hvilke effekter der opnås.

Procesevaluering præsenteres lidt forskelligt af forskellige forfattere. Ifølge Vedung (1998: 166) inkluderer procesevaluering en afdækning af alle typer af effekter herunder intenderede effekter, nuleffekter og bieffekter – forudsete såvel som ikke forudsete, ønskede såvel som uønskede. Heroverfor reserverer Olsen og Rieper (2004: 19) begrebet procesevaluering til procesanalyse og anfører i forlængelse heraf, at det er en svaghed ved procesevaluering, at denne ikke omfatter analyse af effekter.

Tilrettelægges procesevaluering som følgeforskning, hvor evalutor så at sige står på sidelinjen i hele den periode, hvor en indsats følges, og dens effekter vurderes, får procesevaluering visse fællestræk med visse typer af forløbsundersøgelser. Procesevaluering kan baseres både på kvalitativ og kvantitativ metode. Tilrettelægges procesevaluering ex post, må evaluator »optrævle« den historiske proces omkring indsatsen, hvilket kan skabe metodeudfordringer knyttet til aktørernes efterrationalisering af forløbet. Når procesevaluering placeres lavt i evidenshierarkiets optik og lavere end forløbsstudier, er årsagen, at der i procesevalueringsdesignet ikke arbejdes med statistisk kontrol.

Aktionsforskning og formativ evaluering er kendetegnet ved, at der er en dialog mellem forsker og de praktikere, der medvirker i forskningen, undervejs i selve forskningsprocessen. »Praktikere« kan være medarbejdere, ledere, frivillige, brugere af ydelser. Formålet er, at praktikere såvel som forskere bidrager til at udvikle og ændre den givne indsats, ydelse, institution eller organisation.

Der er ingen klar grænse mellem aktionsforskning og formativ evaluering, men de er udviklet i hver deres forsknings- og praksistradition. Aktionsforskningen var oprindeligt udsprunget af organisations- og ledelsesforskning i England i 1950'erne, og blev taget op af den kritiske samfundsforskning i 1970'erne i Skandinavien (Gustavsen 2001).

Formativ evaluering har sin rod i evalueringsforskningen i USA og er knyttet til samarbejde mellem evaluator og praktikere for løbende at forbedre et bestemt program eller indsats.

I praksis forbindes aktionsforskningen med et kraftigere engagement fra forskerens side i de praktiske forbedringer, men også i evalueringsforskningen har forskerens/evaluators rolle et betydeligt konsultativt islæt. Aktionsforskning og formativ evaluering er design, som er betydelig krævende, og som trækker langt flere ressourcer, end forskeren ofte regner med. Der skal investeres megen arbejdstid i at opbygge relationerne til praksisfeltet, og den løbende kontakt mellem praktiker og forsker er en tidsluger. Der er imidlertid store muligheder for læring for både praktiker og forsker. Praktikere får mulighed for større (selv-) erkendelse af deres handlinger og disses konsekvenser, og forskere får et dybere forhold til praksisfeltet, som vanskeligt kan opnås ved den gængse dataindsamling.

Vanskelighederne ved disse design er især knyttet til to forhold. Forskerne kan blive så involveret i praksisfeltet, at de bliver blinde for svagheder i praksis. De bliver alt for indfødte (»go native«). En anden vanskelighed er, at et stærkt praksisfelt kan (mis-)bruge forskningen. Forskeren afkræves en betydelig robusthed i forhold til forskellige interesser og må fastholde sin primære rolle som forsker og samtidig forvalte flere biroller (Launsø og Rieper 2005).

Et eksempel på en procesevaluering på det medicinske område

Et behandlings- og forskningsprojekt i Scleroseforeningen:
Scleroseforeningen gennemfører 2004-2010 et behandlings- og forskningsprojekt, der har til formål at udvikle og afprøve en model for samarbejde mellem etablerede og alternative behandlere for at forbedre behandlingsresultater for mennesker med MS (MmMS). Det er ofte et problem, at de modtager mange forskellige symptomrettede behandlinger, som ikke er koordineret i en behandlingsplan.

Projektet er tilrettelagt som en formativ evaluering. Nogle hundrede patienter får over en årrække tilbudt behandlinger af et team af behandlere, som giver både etablerede og alternative behandlinger. Behandlingsforløbet beskrives gennem interview med behandlere og patienter, og effekten af behandlingerne registreres gennem behandlernes observationer og vurdering og patienternes egen vurdering. Den ansvarlige forsker på projektet tilrettelægger og deltager i behandlerteamets jævnlige møder, hvor behandlerne udveksler erfaringer, og delresultater præsenteres og drøftes, ligesom forskeren er ansvarlig for, at projektets videnskabelige resultater afrapporteres. Den ansvarlige forsker har således både en rolle som facilitator af den tværgående kommunikation i behandlerteamet, opgaven med at udvikle og tilpasse forskningsdesignet samt rollen som analytiker og formidler (Launsø og Haahr 2007).

Et eksempel på en procesevaluering på det sociale område

Kolonien Filadelfia ønskede en evaluering af deres nye indsats i genoptræningscentret Kurhus, hvor man havde valgt at indføre en ny holistisk behandlingsmodel til rehabilitering af mennesker med traumatiske hjerneskader.

Evalueringen havde fire mål: 1) at give rehabiliteringscentret, Kurhus, tilbagemeldinger på deres arbejde med en helhedsorienteret rehabiliteringsmodel undervejs i en 2-årig evalueringsproces, 2) at pege på justeringer og udviklingsmuligheder i rehabiliteringscentrets organisation, 3) at delagtiggøre en bredere offentlighed i de problemstillinger, der knytter sig til implementering af den holistiske og tværfaglige rehabiliteringsmodel, og 4) at sætte modellen ind i den internationale diskussion omkring behandling af traumatiske hjerneskader.

Evalueringen indsamlede primært data ved hjælp af flere på hinanden følgende spørgeskemaer (fire forskellige skemaer) til professionelle, klienter og pårørende gennem en periode på fire år. Spørgeskemaerne blev uddelt ved starten af opholdet på genoptræningscentret, ved afslutningen, samt et år efter opholdet for dem, der nåede så langt i evalueringsperioden. Der blev endvidere indsamlet data ved hjælp af fokuserede gruppeinterview med repræsentanter for de enkelte faggrupper i genoptræningscentret i 2001. Evaluator gennemgik desuden forundersøgelser og klientplaner for

de enkelte klienter. Endelig blev der gennemført et litteraturstudie af den relevante danske og internationale litteratur med fokus på muligheden for reintegration i sociale fællesskaber. Litteraturstudiet dannede efterfølgende grundlag for den internationale perspektivering af rehabiliteringsmodellen. (Høgsbro 2002).

3.6 Kvalitativt casestudiedesign og etnografisk feltstudie

Endnu et trin nede i evidenshierarkiet placeres kvalitative casestudiedesign og etnografiske feltstudier. Kvalitative metodiske design sætter fokus på de involverede aktørers (fx behandlernes, klienternes, beslutningstagernes) oplevelser af en indsats, på deres egen praksis i forhold til denne samt på deres værdier og meningsskabelse i modsætning til at måle resultater på på forhånd definerede effektdimensioner. Kvalitative design kan benytte en vifte af metoder fx interview, deltagerobservation, diskurs- og tekstanalyse, biografier og narrativer. Kvalitative metodiske design vil ofte indgå som en del af procesevaluering.

Anvendes det kvalitative casestudiedesign udvælges et eller flere cases til analyse i den population, der er af interesse. Det kan fx være et udvalg af behandlere eller et udvalg af kommuner, der tilbyder en given type af behandling. Udvælgelsen af case kan baseres på forskellige kriterier. Et case kan fx udvælges som kritisk i forhold til et sæt af antagelser, som unikt/ekstremt eller som fænomenafslørende (Andersen 1997). Kriterierne for udvælgelse af case(s) har stor betydning for, i hvilket omfang det er muligt at generalisere på basis af caseanalysens resultater.

Når det kvalitative casestudie placeres lavt i evidenshierarkiets optik, er årsagen kombinationen af kvalitativ metodik med den risiko for subjektivitet, som denne tilskrives, og casestudiedesignet, med de begrænsninger i generaliseringsmuligheder, som dette tilskrives. Som vi senere skal vende tilbage til er den medicinske del af evidensbevægelsen blevet kritiseret kraftigt for ikke at anerkende kvalitative design. Kritikken har især været rejst fra det psykoterapeutiske felt (Ekeland 2004) samt fra forskere med en humanistisk eller samfundsvidenskabelig baggrund (Launsø og Gannik

2000). Der er initiativer i gang på det medicinske område især inden for medicinsk teknologivurdering for at integrere mere kvalitative indsigter i de medicinske evalueringer (Harden et al. 2004). Dette arbejde må dog stadig betegnes som at være på et forholdsvist indledende stadie.

Et eksempel på et kvalitativt casestudiedesign på det sociale område
Etnografisk undersøgelse af udsatte gruppers livsverden:
Den etnografiske undersøgelse var et element i en større undersøgelse om tilbuddene til mennesker med hjemløshed, misbrug eller sindslidelser som et problem. Undersøgelsen var finansieret af Socialministeriet. Formålet med en etnografiske undersøgelsen var at belyse samspillet mellem dissse menneskers livsverden, forstået som deres forhold til deres personlige livshistorie og de sociale sammenhænge, de færdes i, og det samlede offentlige tilbud til disse grupper. Undersøgelsen bygger på et feltarbejde i fem forskellige regioner i Danmark over 12 måneder, hvor forskerne har tilbragt 1,5 månder i hvert af områderne. En af hovedkonklusionerne er, at der generelt mangler en helhedsforståelse for tilbuddenes sammenhæng og funktion i forhold til den situation, den enkelte bruger befinder sig i. Og der mangler forståelse for, hvordan brugerne støttes i en personlig udvikling, der kan forandre deres situation (Høgsbro m.fl. 2003).

3.7 Erfaringer og eksempler fra praksis

Erfaringsbaseret viden kan genereres ved, at de praksisprofessionelle selv systematisk over tid indsamler og analyserer data vedrørende deres indsats og dennes resultater. Der kan være tale dels om datagenerering, der siger noget om indsats og resultater generelt, dels om beskrivelser af eksempler på god praksis. Tilvejebringelsen af erfaringsbaseret viden kan organiseres enten som en monitorering af en konkret organisatorisk praksis eller som en netværksaktivitet, hvor flere beslægtede organisationer indsamler erfaringer og sammenligner sig på tværs. Erfaringsbaseret viden og eksempler på god praksis betragtes som placeret relativt langt nede i evidenshierarkiet.

Et eksempel på netværk om kvalitetsudvikling i regioner og kommuner

FOKUS blev dannet i 1994 som et forum for udveksling af erfaringer og viden om kvalitetsudvikling i regioner og kommuner og har siden udviklet sig til et netværk, der i dag består af ca. 1850 medlemmer – fortrinsvis medarbejdere og ledere fra bl.a. offentlige institutioner, forskningsverdenen og interesseorganisationer. FOKUS har udgivet godt 100 publikationer og afholdt ca. 90 gåhjemmøder, hvor mange forskellige emner er blevet debatteret. Oplægsholderne og forfatterne er ofte praktikere med særlige erfaringer. FOKUS er finansieret af Det kommunale Momsfond og ledes af en bestyrelse med repræsentanter fra bl.a. AKF, Anvendt KommunalForskning, Danske Regioner, DSI – Dansk Sundhedsinstitut, Kommunernes Landsforening og Kommunale Tjenestemænd og Overenskomstansatte (KTO).

Et eksempel på erfaringsbaseret viden på uddannelsesområdet

En række uddannelsesinstitutioner på ungdomsuddannelsesområdet har etableret et kvalitetsnetværk. På basis af formulering af målbare definitioner af, hvad god kvalitet er, gennemføres en række undersøgelser fx af elevtilfredshed, medarbejdertilfredshed og virksomhedstilfredshed. Undersøgelserne gør det muligt for den enkelte uddannelsesinstitution at sammenligne sig med de øvrige. Herudover benyttes undersøgelserne til at synliggøre god praksis, fx i form af udpegning af årets kvalitetsskole. Endelig benyttes netværket også som afsæt for coaching. De deltagende skoler har mulighed for at benytte sig af en såkaldt besøgsgruppe, hvor konsulenter fra andre skoler med afsæt i identifikation af skolens styrker og svagheder i kvalitetsarbejdet giver input til videre udvikling. For yderligere information se: www.uddannelsesbenchmark.dk

3.8 Eksterne ekspertvurderinger

Nederst i evidenshierarkiet rangeres design, der isoleret tager udgangspunkt i enkeltaktørers vurderinger af effekter. I evalueringslitteraturen benyttes betegnelsen skyggekontrol for denne type af design (Vedung

1998: 163). Design, der fokuserer på eksperters vurderinger, rangeres højere end design, der fokuserer på brugeres vurderinger.

I evidenshierarkiet rangeres ekspertvurderinger således på trinnet under erfaringer og eksempler på god praksis. Ræsonnementet bag dette er, ekspertvurderinger betragtes som baseret på selektive (inden for medicinen kaldet kliniske) erfaringer, som værende subjektive og ikke involverende en systematisk indsamling og analyse af data. Ekspertvurdering, som også diskuteres under betegnelserne kollegial evaluering, fagfællevurdering og peer review, kan imidlertid være andet end enkeltstående ekspertudsagn. Også generering af ekspertviden kan organiseres systematisk. Et eksempel på dette er de såkaldte konsensuskonferencer, der sigter mod at afklare, i hvilket omfang flere eksperter har fælles erfaringer. Et andet eksempel er akkreditering, hvor eksterne eksperter vurderer, om en indsats eller en organisation lever op til et sæt af kriterier og på denne baggrund tager stilling til, om organisationen er kompetent til at udføre sine opgaver.

Rangordningen af ekspertudsagn er vigtig at bemærke, eftersom den medicinske videnskab inden udbredelsen af det randomiserede kontrollerede forsøg og inden evidensbevægelsens indtog typisk baserede sig på medicinske autoriteters udsagn. Kritikere af evidenshierarkiet argumenterer da også for, at evidensbevægelsen tenderer til at underbetone de kvaliteter, der er knyttet til eksperters intuitive forståelse af en indsatssituation eller indsatstype. Denne form for kritik kommer blandt andet til udtryk fra eksperterne selv (se fx Pedersen 2004). Herudover kan den udledes af læringsteoriens såkaldte Dreyfus-model, der betragter eksperter eller virtuoser, som mennesker, der behersker helhedspræget problemløsning, idet de intuitivt, holistisk og synkront kan identificere problem, mål, plan, beslutning og handling (se fx Flyvbjerg 1993: kapitel 2).

Et eksempel på ekspertvurderinger på det medicinske område

Institut for Kvalitet og Akkreditering i Sundhedsvæsenet (IKAS, www.kvalitetsinstitut.dk) arbejder med at udvikle et kvalitets- og akkrediteringssystem kaldet Den Danske Kvalitetsmodel (DDKM) for det danske sundhedsvæsen. Formålet med DDKM er: 1) at udvikle et evalueringsgrundlag i form af standarder med tilhørende indikatorer, 2) at fremme kontinuerlig klinisk, faglig og organisatorisk kvalitetsforbedring af pati-

entforløbene, 3) at gennemføre en ekstern vurdering og akkreditering af de involverede institutioner og 4) at understøtte gennemsigtighed og gennemskuelighed af kvaliteten i sundhedsvæsenet. Til at gennemføre de eksterne vurderinger vil der blive uddannet team af seniore medarbejdere fra relevante sundhedsprofessioner.

Et eksempel på ekspertvurderinger på det sociale område

I en undersøgelse af sociale indsatser over for udsatte grupper blev anvendt en særlig høringsmetode, kaldet audit, hvor der benyttes høringsforløb tilrettelagt for at få præsenteret professionelles syn på indsatsen og med vægt på især vurdering af, hvilke behov der bliver dækket, og hvilke der ikke bliver, samt en indkredsning af eventuelle barrierer for udførelse af godt arbejde og endelig, hvilke succeser og hvilke fiaskoer der ses.

De professionelle paneldeltagere har været psykologer, læger, sygeplejersker, socialrådgivere, socialpædagoger, håndværksuddannede og personer med helt andre faglige baggrunde. Deres ansættelsesmæssige baggrund har været kommunale socialforvaltninger, distriktspsykiatri, alkoholbehandlingsinstitutioner, narkobehandlingsinstitutioner, væresteder, § 94-botilbud, kriminalforsorg, sygeafdelinger, socialpsykiatriske projekter samt private organisationer og foreninger. Nogle af paneldeltagerne arbejder i blivende og tunge institutionelle tilbud andre i små måske midlertidige puljefinansierede tilbud.

Ud fra cases indsamlet i hver af de regioner, der indgik i undersøgelsen, er der udvalgt en række beretninger – enten længere dele af camouflerede livshistorier eller kortere dele, hvor særlige begivenheder indgår.

Selve auditdagene har været bygget op med to panelrunder, gennemført henholdsvis formiddag og eftermiddag, med forskelligt panel, hvor hver paneldeltager har givet et bud på 3-4 aktuelle beretninger fra regionen, efterfulgt af spørgsmål fra den resterende del af panelet, og yderligere efterfulgt af diskussion og spørgsmål fra salen, hvor der har været mellem 10 og 15 personer. Hele forløbet er blevet optaget på bånd, der efterfølgende er udskrevet i sin helhed (Brandt og Kirk 2003).

3.9 Brugervurderinger

Undersøgelser af brugernes (patienternes, klienternes) oplevelse af interventioner betragtes af evidenshierarkiets fortalere som havende meget begrænset evidensstyrke. Argumentet er, at brugerne alene af den grund, at de oplever, at deres problemer tages alvorligt, kan vurdere indsatser positivt på trods af, at disses effekt er nul eller måske endog direkte skadelig. Som illustration af denne problematik henviser evidensbevægelsens fortalere ofte til visse typer af kriminalpræventive interventioner rettet mod unge. På dette felt har lodtrækningsforsøg nemlig vist, at indsatser, som de unge vurderede som positive, på sigt har bidraget til at forøge de unges kriminalitet snarere end at reducere denne, formodentlig fordi indsatserne har skabt negative rollemodeller for de unge ved enten at bringe de unge sammen i grupper eller konfrontere dem med ældre kriminelle. Det gælder mere generelt, at der er en tendens til, at undersøgelsesdesign, der ikke baserer sig på RCT viser mere positive effekter af indsatsen end studier, der er baseret på RCT (Oakley 2000, ss 251ff).

Også diskussionen om brugerundersøgelser er imidlertid mere nuanceret end som så. For det første kan der argumenteres for, at brugerundersøgelser har et potentiale til at bidrage til at forøge effekten af visse typer af indsatser. For eksempel kan undersøgelser af brugere (herunder frafaldne potentielle brugere) give viden om barrierer for deltagelse i interventioner fx i forbindelse med forebyggende indsats og tidlige diagnostik, viden, der kan anvendes til at forøge deltagelsen og dermed effekten af den forebyggende indsats. For det andet kan der peges på faser i interventioner og på typer af interventioner, hvor samspillet mellem bruger og behandler er så intensivt, at brugeroplevelsen må formodes at have betydning for effekten. På det medicinske område kan der fx ikke behandles, uden at der er stillet en diagnose, og en fejldiagnose kan have alvorlige følger. Visse typer af diagnoser kan ikke stilles uden et tæt samspil mellem patient og behandler, et samspil, hvor patientens oplevelse af situationen formodentlig influerer på diagnosticeringen. Ligeledes er der inden for psykoterapi, psykologi og pædagogik blevet stillet spørgsmåls ved evidensbevægelsens forestilling om den kontekstfrie intervention, og det er blevet understreget, at både terapeutens og klientens tro på metoden er vigtigere end metoden som sådan (Wampold 2001, Ekeland, 2004). Inden for denne forestilling om

kontekstuelle indsatser bliver brugerundersøgelser såvel som ekspertvurderinger centrale kilder til viden om effekt.

I litteraturen om brugervurderinger sondres der ofte mellem brugervurderinger knyttet til præstationsmålinger og igangsat som led i oppefra og ned organiserede styringsinitiativer, også benævnt brugertilfredshedsundersøgelser, og mere dialogisk tilrettelagte brugervurderinger igangsat med afsæt i et deliberativt demokratisk argument (Andersen, 2003; Hansen, 2003). Såvel den såkaldte KUBI-model udviklet af den almennyttige forening Socialt Udviklingscenter Storkøbenhavn (www.sus.dk), og BIKVA-modellen er eksempler på konkrete dialogiske modeller, der sigter mod at inddrage brugerne i kvalitetsvurdering og på basis heraf udvikle kvaliteten i fagprofessionel praksis (Dahler-Larsen & Krogstrup 2003; Krogstrup 2006).

Et eksempel på brugervurderinger på det medicinske område

Enheden for Brugerundersøgelser, Region Hovedstaden (www.patientoplevelser.dk) har i en årrække gennemført undersøgelser af patientoplevelser i sundhedsvæsenet. Enheden har blandt andet ansvar for gennemførelsen af Den Landsdækkende Undersøgelse af Patientoplevelser (LUP). Formålet med LUP er at sammenligne patientoplevelser på sygehus- og specialeniveau samt sammenligne patientoplevelser over tid. Undersøgelsen sætter fokus blandt andet på kliniske ydelser, patientsikkerhed, medinddragelse og kommunikation, behandlingsforløb, udskrivelse, ventetid og frit sygehusvalg. I den seneste runde af LUP blev data indsamlet via et spørgeskema, der ultimo august 2006 blev sendt til 26.313 patienter, der havde været indlagt på 53 sygehuse i perioden medio marts til medio juni 2006 (Enheden for Brugerundersøgelser, 2006).

Et eksempel på brugervurderinger på det sociale område

Med ikrafttrædelsen af Lov om Social Service i 1998 blev det traditionelle institutionsbegreb for voksne handicappede ophævet og erstattet med et boligbegreb. Samtidig indførtes en skelnen mellem botilbud og serviceydelser. I forlængelse af loven blev der i 2001-2002 gennemført en evaluering, der havde til formål at afdække, hvor langt amter og kommuner var kommet i omstillingen fra institution til individuelle botilbud. Metodisk

blev evalueringen i betydeligt omfang baseret på brugervurderinger, idet brugerbegrebet inkluderede såvel de handicappede som deres pårørende. Datamaterialet udgjordes af allerede gennemførte KUBI-evalueringer af botilbud, supplerende KUBI-evalueringer samt fokusgruppeinterviews (Perit m.fl. 2002).

3.10 Opsummering

Som det er fremgået, rangordner fortalerne for evidenshierarkiet undersøgelsesdesign i op til 9 niveauer. Det randomiserede, kontrollerede eksperiment betragtes klart som det undersøgelsesdesign, der producerer viden af størst validitet, mens kvalitative design og aktørmodeller (ekspertvurderinger og brugervurderinger) betragtes som placeret i bunden af hierarkiet. Mellem disse yderpunkter rangeres en række design frem for alt ikke-randomiserede eksperimenter og forløbsstudier, herunder kohorteundersøgelser. På trods af at evidenshierarkiet indeholder op til 9 niveauer, er evidensbevægelsens betoning af RCT ofte så markant, at det næsten udelukkende er denne guldstandard, der associeres med hierarkiet. Derfor vil der i det følgende blive lagt hovedvægt på den debat og de problematikker, der knytter sig til anvendelsen af dette forskningsdesign.

4 Evidenshierarkiet, som det fremgår af evidensorganisationernes egne vejledninger

De evidensproducerende organisationer udarbejder og henviser til vejledninger i hvordan systematiske forskningsoversigter skal udarbejdes, herunder efter hvilke kvalitetskriterier de primære studier udvælges, hvordan disse skal kvalitetsvurderes, og hvordan deres resultater skal syntetiseres. På grundlag af en gennemgang af en række evidensorganisationers hjemmesider er vejledninger (guidelines, handbooks mv.) blevet identificeret og deres anbefalinger systematiseret, se bilag 1.

Følgende organisationers hjemmesider er blevet gennemgået:

- Cochrane Collaboration, international evidensproducerende organisation på det medicinske område.
- The Nordic Cochrane Centre, regionalt center inden for Cochrane Collaboration.
- Centre for Reviews and Dissemination (CRD), engelsk evidensproducerende organisation placeret på York's universitet med betydelig finansiering fra det engelske sundhedsministerium.
- National Institure for Health and Clinical Excellence (NICE), engelsk evidensproducerende organisation på sundhedsområdet finansieret af det engelske sundhedsministerium.
- The Campbell Collaboration (C2), international evidensproducerende organisation på områderne socialt arbejde, uddannelse og kriminologi.
- Nordisk Campbell Center (NC2), regionalt center inden for Campbell Collaboration.

- Institutet för utveckling av metoder i socialt arbete (IMS), enhed under den svenske Socialstyrelse. IMS er den svenske kontakt til Campbell samarbejdet.
- Social Care Institute for Excellence (SCIE), engelsk evidensproducerende organisation på området socialt arbejde med betydelig finansiering fra det engelske sundhedsministerium.
- The Evidence for Policy and Practice Information and Coordinating Centre (EPPI), evidensproducerende enhed på uddannelsesområdet etableret i regi af Social Science Research Unit på University of London.
- What Works Clearinghouse (WWC), enhed under det amerikanske undervisningsministeriums forskningsenhed »Institute of Education Sciences«. WWC indgår i en kontrakt med bl.a. Campbell Collaboration.

I det følgende redegøres der for, hvilken politik disse organisationer har, hvad angår udarbejdelsen af systematiske reviews, herunder hvor de lægger snittet for, hvilke design de inkluderer, hvordan de kvalitetsvurderer primærstudier, samt hvordan de syntetiserer resultaterne.

4.1 Snittet for inklusion

Af de 10 organisationer er der seks organisationer, der eksplicit i egne retningslinjer angiver, at de arbejder ud fra evidenshierarkiets logik. Det drejer sig om Cochrane Collaboration, Centre for Reviews and Dissemination (CRD), The Campbell Collaboration (C2), Nordisk Campbell Center (NC2), National Institure for Health and Clinical Excellence (NICE) og What Works Clearinghouse (WWC). Herudover er der to organisationer, der implicit angiver, at de arbejder ud fra evidenshierarkiets logik, idet de henviser til nogle af ovenstående. Det drejer sig om The Nordic Cochrane Centre, der henviser til Cochrane Collaboration, og Institutet för utveckling av metoder i socialt arbete (IMS), der henviser til Campbell Collaboration. Fælles for de nævnte organisationer er, at de eks- eller implicit anfører, at de prioriterer primærstudier, der anvender RCT, højest.

Positionen kan illustreres ved at citere fra Cohrane's håndbog (Higgins & Green, 2006: section 4.2.4):

»Certain study design are superior to others when answering particular questions. Randomised controlled trials (RCT) are considered by many the sine qua non when addressing questions regarding therapeutic efficacy, whereas other study designs are appropriate for addressing other types of questions. For example questions relating to aetiology or risk factors may be addressed by case-control and cohort studies….. Because Cochrane reviews address questions about the effects of healthcare they focus primarily on RCT.«

Til denne hovedregel tilføjes efterfølgende, at det nogle gange kan forsvares at basere systematiske reviews på ikke randomiserede primærstudier. Dette gælder, fx hvis der ikke findes RCT, fordi effekterne af en intervention er så dramatiske (=har så stor effekt), at det er uetisk at gennemføre RCT.

Campbell Collaboration synes at være på samme linje, men deres vejledninger ikke er helt klare. På NC2's hjemmeside anføres det i redegørelsen for, hvordan man udarbejder en protokol, at det i denne skal specificeres, hvilke typer primærstudier der vil blive inkluderet, hvorefter det tilføjes »sædvanligvis RCT«. Andre steder fremgår det, at reviews normalt ikke bør gennemføres, hvis der ikke findes RCT, men samtidig at reviews ikke behøver at være baseret alene på RCT-designede primærstudier. Ikke randomiserede studier kan også inkluderes. På C2's hjemmeside findes en vejledning, som åbner yderligere op. Her anføres det, at Campbell-reviews ud over ovenstående kan inkludere resultater fra en vifte af metoder, kvantitiative såvel som kvalitative. Der kan fx være tale om studier, der belyser implementeringsprocesser, eller studier, der beskriver de subjektive erfaringer hos personer, der har fået tilbudt den pågældende intervention. De lidt uldne vejledninger synes at afspejle, at der løbende er en debat om, hvilke inklusionskriterier der bør anvendes.

I et videnskabsteoretisk perspektiv kan tilgangen hos de evidensorganisationer, der tager afsæt i evidenshierarkiet, karakteriseres som baseret på en empirisk analytisk videnskabsopfattelse, hvor en sand værdi kan måles,

og hvor kausale relationer (at en indsats fører til bestemte resultater) bedst afdækkes ved RCT. Primærstudier, der anvender RCT-design, betragtes som stærkest, andre design som svagere. Resultater fra primærstudier, der benytter kvalitative design, betragtes som kilder, der kan støtte eller underbygge resultaterne fra de eksperimentelle og kvasieksperimentelle design.

Endelig er der to organisationer, der eksplicit angiver, at de ikke tager afsæt i rangordning af design. Det drejer sig om de to engelske organisationer The Evidence for Policy and Practice Information and Coordinating Centre (EPPI) og Social Care Institute for Excellence (SCIE).

EPPI anfører, at de, i modsætning til Cochrane og Campbell, der karakteriseres som udelukkende interesseret i at besvare spørgsmål om, hvilke interventioner der virker, producerer systematiske reviews, der sætter fokus på en bred vifte af problemstillinger. Ud over spørgsmål knyttet til, hvilke interventioner der virker, kan der fx være tale om spørgsmål som, hvordan opleves det at modtage en intervention? Eller hvorfor opstår et givent fænomen? Konsekvensen af at arbejde med forskellige typer af problemstillinger er, at der er behov for at inkludere forskellige typer af evidens. EPPI arbejder derfor også med at integrere resultater fra RCT-design med resultater fra andre typer af design.

SCIE er på samme linje, men åbner op for en endnu bredere vidensbase. I en publikation udgivet af SCIE (Pawson, Boaz, Grayson, Long & Barnes, 2003) understreges det, at opbygningen af en evidensbaseret vidensbase på det sociale område må trække på fem typer af videnskilder: organisatorisk viden, praksisforankret viden, brugerforankret viden, forskningsbaseret viden samt viden genereret i den politiske verden. Det anføres eksplicit, at alle kilder er vigtige, og at listningen af dem ikke er udtryk for et hierarki. Herudover listes også en række mere konkrete potentielle vidensformer: eksperimentelle og kvasieksperimentelle tilgange, monitorering, konsultering (konsensusdannelse baseret blandt andet på forskellige former for dialog), kvalitative casestudier, aktionsforskning, brugervurderinger, procesevaluering, audit og inspektion, høringer, dialogiske tilgange og praksiserfaringer. Det ses heraf, at SCIE definerer de videnskilder, der er relevante for opbygning af en evidensbaseret vidensbase på det sociale område væsentligt bredere end Campbell, og at organisationen er åben over for at inkludere alle de design, der indgår i evidenshierarkiet, ja faktisk lægger

flere til. Mens det er metododologiske kriterier, der er de primære for Campbell, er det relevanskriteriet, der er det primære for SCIE. Som det formuleres af fremtrædende SCIE-folk: »Relevance is the hallmark of social work research, not a particular approach to research methodology« (Marsch and Fisher 2005: 12). Den brede vidensbase reflekteres også i terminologien. Således benytter SCIE konsekvent betegnelsen vidensbidrag og ikke primærstudier, ligesom SCIE's reviews betegnes »knowledge reviews«.

4.2 Kvalitetsvurdering af primærstudier

Også i forhold til kriterier for kvalitetsvurdering af primærstudier er der visse forskelle i terminologi og anbefalinger organisationerne imellem. Anbefalingerne hos Cochrane Collaboration (inklusive det nordiske center), Campbell Collaboration (ligeledes inklusive det nordiske center) og CRD er i høj grad overensstemmende, ligesom de krydshenviser til hinanden. Cochrane Collaboration og CRD er de organisationer, der har udgivet de mest omfattende manualer for udarbejdelse af systematiske oversigter.

I Cochrane Handbook nævnes fire mulige kilder til bias (fejl), som alle primærstudier bør kvalitetsvurderes i forhold til. Det drejer sig om: 1) Selektionsbias, 2) Performancebias, 3) Attritionsbias og 4) Detektionsbias. Selektionsbias er som tidligere nævnt defineret som bias, der opstår i allokeringen af forsøgspersoner til den undersøgte indsats. Performancebias er defineret som bias, der opstår i forsøgsperioden i leveringen og modtagelsen af indsatsen. Performancebias kan fx opstå, hvis personer i kontrolgruppen finder ud af, at de indgår i kontrolgruppen, og derfor på egen hånd opsøger andre indsatser eller måske prøver at gøre det samme som deltagerne i forsøgsgruppen. Ligeledes kan de opstå, hvis behandlere i kontrolgruppen, der er forudsat at levere praksis as usual, udvikler deres praksis fx ved at indlåne elementer fra den praksis, der tilbydes forsøgsgruppen. De nævnte problemer diskuteres til tider under betegnelsen imitationsproblemet (Vedung, 1998: 142). Attritionsbias er defineret som bias, der opstår som følge af, at frafaldet er forskelligt i henholdsvis forsøgs- og kontrolgruppen. Endelig er detektionsbias defineret som bias, der opstår, hvis ef-

fekten på deltagerne i henholdsvis forsøgs- og kontrolgruppen måles på forskellig vis. Konsekvensen af bias er, at forsøgs- og kontrolgruppen reelt ikke bliver sammenlignbare.

Med henblik på at undgå attritionsbias anbefales det ofte at anvende den såkaldte »intention to treat« (ITT)-analyse. ITT-analysen indebærer, at alle de personer, som det var hensigten at behandle, inkluderes i effektanalysen. Kan man ikke få fat i de personer, der er faldet fra undervejs, antages det, at der ikke har været nogen effekt for de pågældende. ITT-analysen kritiseres til tider for at undervurdere effekten og kritikere anbefaler i stedet at anvende den såkaldte »totally treated« (TOT)-analyse. TOT-analysen inkluderer alene de personer, der har gennemført behandling. Og kritikere af TOT-argumenterer modsat, at TOT-analysen tenderer til at overvurdere effekten, fordi de, der faldt fra undervejs, formodentlig faldt fra på grund af manglende effekt.

CRD har en overlappende, men lidt bredere tilgang til kvalitetsvurdering af primærstudier. På CRD's hjemmeside[3] findes en manual til udarbejdelse af systematiske oversigter kaldet »Undertaking Systematic Reviews of Research on Effectiveness«[4]. Manualen er tænkt som en støtte til arbejde i CRD's indsatsområder, navnlig sundhedsområdet og dele af socialområdet.

Ifølge manualen bliver kvalitetsvurdering af studier diskuteret inden for følgende terminologi:

- Studiekvalitet (metodologisk kvalitet)
  Graden af, hvor godt et studie inkorporerer tiltag, der minimerer bias og dermed styrker den interne validitet. Synes at være et overordnet kvalitetskriterium.
- Bias (systematiske fejl)
  Tendensen til at producere resultater, der systematisk afviger fra det »sande« resultat. Ikke-biased resultater er internt valide. Biastyper: Selektionsbias/Performance bias/Measurement bias (kaldes »detection bias« hos Cochrane)/Attrition Bias.
- Intern validitet
  Hvor sandsynligt det er, at et studies resultater er tæt på »sandheden«. Intern validitet er forudsætningen for ekstern validitet.
- Ekstern validitet

Graden af, hvor anvendelige/overførbare de observerede effekter er uden for det studerede område.

I rapporten nævnes også en række metoder til beskyttelse mod bias i primærstudier. Disse er i forlængelse af det foregående ikke overraskende: randomisering med skjult allokation, blinding og anvendelse af »intention to treat analysis«.

En mere fyldig redegørelse for vurderingen af metodologisk kvalitet i primærstudier findes i Farrington (2003). David P. Farrington er formand for Campbell Collaboration Crime and Justice Group, og det må antages, at han afspejler et gennemgående synspunkt i Campbell Collaboration som helhed. Farrington nævner fire metodologiske kvalitetskriterier, som hjørnestene i Campbells tilgang gennem årene:

- Statistisk konklusionsvaliditet
  Omhandler, hvorvidt den formodede årsag (interventionen) og den formodede effekt (outcome) er sammenhængende. Mål for effektstørrelse og tilhørende konfidensintervaller bør beregnes. Sandsynligheden for at få den observerede effekt under antagelse af nulhypotesen bør også beregnes.
  Største trusler er:
  – Utilstrækkelig statistisk styrke, fx ved for lille sample-størrelse.
  – Anvendelse af upassende statistiske test.
  – Heterogenitet på tværs af sammenligningsgrupperne.
- Intern validitet
  Refererer til kausalitetsspørgsmålet om, hvorvidt interventionen virkelig forårsagede ændringer i outcome.
- Konstruktionsvaliditet
  Konstruktionsvaliditet refererer til, om den operationelle definition er tilstrækkelig for målingerne af den teoretiske model, der ligger til grund for interventionen og outcomemålene. Altså om man måler det, man ønsker at måle.
  Største trussel er:
  – Om interventionen gennemføres efter hensigt og plan.
  – Om outcomemålene er troværdige, dvs. gode indikatorer for de outcome, der ønskes.

- Ekstern validitet
  Refererer til generalisérbarheden af de kausale sammenhænge til andre situationer end dem, det pågældende studie vedrører.

WWC har udarbejdet sit eget kategoriseringssystem for kvalitetsvurdering af primærstudier. Systemet, hvis kvalitetskriterier i høj grad er overlappende med Cochranes kriterier for vurdering af risiko for bias, vil af praktiske grunde blive behandlet i kapitel 5.

De to tidligere nævnte organisationer, der ikke tager afsæt i evidenshierarkiets rangordning af design, EPPI og SCIE, arbejder i forlængelse heraf naturligt nok med et bredere sæt af kriterier for kvalitetsvurdering af primærstudier.

EPPI anfører i sin vejledning, at kvaliteten af primærstudier bør baseres på fire kriterier: 1) metodologisk kvalitet, 2) metodologisk relevans, 3) emnerelevans og 4) overordnet vurdering. Med metodologisk kvalitet menes en vurdering af pålideligheden af primærstudiets resultater set i lyset af de alment accepterede normer for gennemførelse af netop den specifikke type af design, der er anvendt. Med metodologisk relevans menes en vurdering af hensigtsmæssigheden af det anvendte design set i lyset af den problemstilling, der er i fokus i det systematiske review. Med emnerelevans menes en vurdering af hensigtsmæssigheden af primærstudiets fokus set i lyset af den problemstilling, der er i fokus i det systematiske review. Endelig menes med overordnet vurdering en vægtning af de allerede nævnte delvurderinger, hvorved der frembringes en samlet vurdering af primærstudiets vidensbidrag til det systematiske review.

SCIE (Pawson, Boaz, Grayson, Long & Barnes, 2003) anbefaler, at kvalitetsvurdering af vidensbidrag baseres på de såkaldte TUPURAS-kriterier: 1) Gennemsigtighed (transperency), 2) præcision (acuracy), 3) formålsrettethed (purposivity), 4) nytte (utility), 5) forsvarlighed (propriety), 6) tilgængelighed (accessibility) og 7) specificitet (specificity).

Følgende spørgsmål må med afsæt i disse kriterier besvares i forhold til hvert enkelt vidensbidrag:

1. Er grunden for den specifikke viden og baggrunden for den klar?
2. Er bidraget ærligt velfunderet på relevant viden?
3. Er de anvendte metoder egnet til formålet?

4. Er det anvendeligt? Besvares de spørgsmål, der stilles?
5. Er det lovligt og etisk forsvarligt?
6. Er det almenforståeligt?
7. Lever det op til de standarder, der gælder for netop denne type viden?

Kriterierne 1 til 6 er generiske, mens kriterium 7 understreger vigtigheden af også at vurdere de enkelte bidrag på deres egne præmisser. TUPURAS-kriterium 7 har således fællestræk med EPPI's kriterium 1, men er på grund af SCIE's åbning mod at inkludere også organisatorisk viden og viden genereret i den politiske verden ikke udelukkende knyttet til metodologisk kvalitet.

Anbefalingerne fra EPPI og SCIE illustrerer, at evidensproducerende organisationer, der ikke tager afsæt i evidenshierarkiet, overordnet arbejder med to typer af kriterier. For det første vurderes kvaliteten af primærstudierne på deres egne betingelser, betingelser, der varierer mellem design og vidensdomæner. For det andet vurderes primærstudierne i forhold til deres relevans, metodologisk såvel som emnemæssigt i forhold til den problemstilling, der er i fokus i det systematiske review.

4.3 Syntetisering af resultater

De evidensproducerende organisationer, der prioriterer primærstudier med RCT-design, betragter i forlængelse heraf metaanalyse som den ideelle syntetiseringsmetode. I dens simpleste form udøves metaanalyse i to trin. Først beregnes standardiserede effektmål for de enkelte primærstudier ved at sammenligne interventions- og kontrolgruppen. Herefter summeres dataene på tværs af primærstudierne i et overordnet effektmål. Cochrane Collaboration har udviklet sin egen softwarepakke, kaldet Cochrane RevMan, til udarbejdelse af reviews inkluderende hjælp til gennemførelsen af metaanalyse.

Hvis der ikke foreligger tilstrækkeligt datamateriale af god kvalitet fra RCT-designede primærstudier, kan data fra andre typer af design som nævnt inkluderes. I disse tilfælde anbefales det at syntetisere data fra forskellige typer af design separat. Muliggør data statistisk syntetisering anbefales dette frem for narrativ syntetisering.

De organisationer, der ikke tager afsæt i rangordning af design, har en mere pluralistisk tilgang til syntetisering. EPPI anfører, at den valgte synteseform varierer mellem forskellige reviews i forhold til, hvilken problemstilling der er i fokus, og hvilke design der har været anvendt i de primærstudier, hvis resultater skal syntetiseres. Der sondres mellem syntetiseringsformerne metaanalyse, narrativ syntetisering og konceptuel syntetisering. Ligesom hos Cochrane og Campbell anbefales metaanalyse i reviews, der har fokus på, om givne interventioner og indsatser virker, og hvor der benyttes data fra eksperimentelle design. Narrativ syntetisering anbefales, når data fra en bredere vifte af design skal syntetiseres. Conceptuel syntetisering anbefales i systematiske reviews, der har fokus på at forstå givne fænomener, og hvor flere konceptuelle perspektiver bringes i spil med sigte på begrebsudvikling. I nogle reviews kombineres flere syntetiseringsformer.

5 Evidenshierarkiet, som det praktiseres ved udarbejdelsen af systematiske forskningsoversigter

I det foregående var fokus på evidensorganisationernes politikker og vejledninger for udarbejdelse af systematiske reviews. I dette kapitel vil vi se på, hvordan deres praksis er, herunder om der er overensstemmelse mellem politik og praksis. Uanset hvilken politik der er vedtaget vedrørende principiel rangorden af forskningsdesign, er spørgsmålet, hvilke typer af viden man vil acceptere som gyldig evidens i en konkret situation. Det første afsnit (5.1) indeholder en række eksempler på forskningsoversigter. Vi fremhæver, hvor snittet er blevet lagt, og efter hvilke kriterier vurderingen af primærstudier er foregået ifølge beskrivelsen i forskningsoversigterne. De følgende afsnit, snittet for inklusion (5.2), kvalitetsvurdering af primærstudier (5.3) og syntetiseringsmåder (5.4) omfatter dels en opsummering fra eksemplerne i afsnit 5.1, dels supplerer vi med mere overordnet viden om praksis genereret via egne opgørelser og andre kilder.

5.1 Eksempler på systematiske forskningsoversigter

Lad os for at sætte lidt kød på, hvordan de forskellige typer af forskningsoversigter kan se ud, give et par eksempler.

Eksempler fra sundhedsområdet

Figur 5.1 viser hovedelementerne i en forskningsoversigt. Eksemplet vedrører, hvorvidt indsatser med antibiotisk medicin har effekt på patienter med skrumpelever og blødende mavesår. Oversigten inkluderer resultater fra 15 analyser med baggrund i 11 gennemførte RCT. I vurderingen af primærstudierne blev tre af de 11 RCT vurderet til »A« i betydningen, at risiko for bias blev bedømt som lav, mens 8 studier blev vurderet til »B«, som betød, at risikoen for bias blev bedømt som moderat. Grunden til at nogle studier blev vurderet til »B«, var typisk, at den metode, der havde været anvendt til at allokere personer til henholdsvis forsøgs- og kontrolgruppen, var ufuldstændigt beskrevet. Resultaterne af begge kategorier af primærstudier var syntetiseret ved hjælp af metaanalyse.

Figur 5.1 Et eksempel på et Cochrane review om antibiotikaprofylakse

Faser:	**Eksempel: Antibiotikaprofylakse for cirrosepatienter med gastrointestinal blødning**
Problemstilling:	Virker forebyggende behandling med antibiotika på cirrosepatienter med gastrointestinal blødning?
Søgning:	Der blev fundet 43 publikationer, der potentielt kunne bidrage med evidens. 19 blev umiddelbart ekskluderet, da de alene indeholdt litteraturoversigter.
Kritisk vurdering:	Af de resterende 24 publikationer opfyldte 15, refererende til 11 RCT-forankrede primærstudier kriterierne for inklusion, mens 7 refererende til 4 andre typer af kliniske forsøg ekskluderedes. Endelig blev 2 publikationer lagt til side for efterfølgende vurdering. Af de 11 RCT forankrede primærstudier blev 3 scoret »A«, hvilket betyder, at risikoen for bias vurderes som lav, mens de resterende 8 blevet scoret »B«, hvilket betyder, at risikoen for bias vurderes som moderat. Data vedrørende design, deltagere, interventionstyper samt resultatmål blev uddraget fra de 11 primærstudier og analyseret statistisk, idet de 11 RCT blev opdelt i to grupper af studier, henholdsvis forsøg, hvor kontrolgruppen blev tilbudt placebo eller slet ikke blev tilbudt behandling, versus forsøg, hvor kontrolgruppen blev tilbudt en anden type af behandling.
Konklusion:	Forebyggende behandling med antibiotika reducerer dødelighed og bakterielle infektioner og bør anbefales.

Bem.: Se også omtalen i Kürstein P., Kjellberg J., Herbild L., Olsen K. R., Willemann M., Søgaard J. & C. Gludd m.fl. (2005: 28-29).

Figur 5.2 vedrører et eksempel omhandlende psykologisk debriefing gennemført med sigte på at forebygge post traumatic stress disorder. Også i denne forskningsoversigt blev alle andre end RCT-baserede primærstudier ekskluderet.

Figur 5.2 Et eksempel på et Cochrane review om psychological debriefing

Faser:	**Psychological debriefing for preventing post traumatic stress disorder (PTSD)**
Problemstiling:	Virker rutinemæssig benyttelse af psykologisk debriefing efter traumatiske oplevelser forebyggende i forhold til udvikling af post traumatic stress disorder?
Søgning:	Der blev foretaget elektronisk søgning i MEDLINE, EMBASE, PsychLit, PILOTS, Biosis, Pascal, Occ. safety and Health, SOCIOFILE, CINAHL, PSYCINFO, PSYNDEX, SIGLE, LILACS, CCTR, CINAHL og NRR og manuel søgning i Journal of Traumatic Stress. Førende forskere blev kontaktet direkte. Søgekriterierne inkluderede randomiserede og kvasirandomiserede primærstudier. 21 studier blev ekskluderet, hovedsageligt fordi de ikke var randomiserede.
Kritisk vurdering:	15 primærstudier opfyldte inklusionskriterierne. Den metodologiske kvalitet var varierende, men hovedparten af primærstudierne scorede svagt. Syntetisering blev gjort via metaanalyse. Data fra 6 primærstudier kunne ikke inkluderes i metaanalysen. Resultaterne fra disse blev sammenfattet i teksten.
Konklusion:	Der er ingen evidens for, at rutinemæssig brug af enkelt sessions debriefing efter traumatiske oplevelser reducerer psychological distress og forebygger udvikling af post traumatic stress disorder (PTSD). Psykologisk debriefing er enten ligestillet med, eller dårligere end, kontrol- eller uddannelsesindsatser rettet mod forebyggelse af PTSD. Der er tegn på, at rutinemæssig debriefing kan forøge risikoen for PTSD og depression.

Bem.: (Cochrane Database of Systematic Reviews 2006, Issue 4; Status:Commented, Copyright © 2006 The Cochrane Collaboration. Published by John Wiley and Sons, Ltd. DOI: 10.1002/14651858.CD000560. First published online by Rose, S.; Bisson, J., Churchill, R. and Wessely, S.: 22 April 2002).

I begge forskningsoversigter fra Cochrane blev de primære RCT-studier, og andre, vurderet. Men vurderingerne blev gjort ud fra forskellige kriterier i de to oversigter. I oversigten vedr. psykologisk debriefing blev primærstudiernes metodologiske kvalitet vurderet uafhængigt af hver af reviewerne, og der blev anvendt tre forskellige vurderingsmetoder med efterfølgende sammenligning af resultaterne. Efterfølgende blev en af skalaerne valgt og studierne rangordnet efter metodologisk kvalitet.

Den første vurderingsmetode var metoden beskrevet I C2's håndbog med skalaen: »adekvat«, (fuldt randomiseret), »intermediate« (ikke fuldt ud randomiseret) og »inadequate« (ingen randomisering). Den anden vurderingsmetode var den såkaldte CCDAN kvalitetsvurderingsskala (Moncrieff et al. 2001). Den tredje var en vurderingsmetode specielt udviklet til vurdering af studier af psykologisk debriefing (Kenardy and Carr 1996). Kvalitetskriterierne er her klar definition og afgrænsning af målgruppe og debriefingmetode, randomisering, anvendelse af både selvrapporteringer og objektive observationer som basis mål, indhentning af outcome mål på

et relevant tidspunkt under hensyntagen til problemer i målgruppen og anvendte debriefingmetoder. Kvalitetsvurderingen af primærstudierne foregik således ganske grundigt ifølge artiklens forfatter (Rose et al. 2002: 6-7). Man kan sige, at kvalitetsvurderingen er blevet »kontekstualiseret«, dvs. skræddersyet til den pågældende indsats (psykologisk debriefing), samtidig som vurderingen fandt sted inden for evidenshierarkiets logik med RCT som den gyldne standard.

Den pågældende pågældende systematiske forskningsoversigt er imidlertid blevet kritiseret for at at anvende for snævre kvalitetskriterier ved vurderingen af primæstudier:

In conflict, following disaster or accident, naturalistic studies, often conducted opportunistically, remain useful and have considerable heuristic value despite methodological shortcomings particularly relating to sample selection and randomisation to different treatment conditions. Applying the stringent criteria demanded by the arbiters of EBM such as the Cochrane library to trials of preventive interventions means that much useful work might go unpublished…RCT have become so divorced from clinical reality that their findings become meaningless…RCT are not the sine qua non of EBM and debriefing studies challenge their hegemony and lend credibility to observational studies. This has important implications for the ways in which the quality and value of research evidence are assessed both in social psychiatry and empirical science in general.« (Extract from Deahl, 2006: 12).

De to ovennævnte eksempler på forskningsoversigter illustrerer, at selv for en bestemt evidensproducerende organisation, som arbejder inden for evidens hierarkiets tankegang, foregår vurderingen af primærstudier på forskellige måder.

Eksempler fra det sociale område og arbejdsmarkedsområdet

Figur 5.4 viser hovedelementerne i en forskningsoversigt fra Campbell om multisystemisk terapi (MST) til behandling af sociale, følelsesmæssige og adfærdsmæssige problemer hos unge i alderen 10-17 år. MST blev udviklet på the Family Services Research Center på det medicinske univer-

sitet i South Carolina. Det er en familiebaseret terapi, som foregår i personens eget hjem eller nærområde. Behandlingsteamet er til rådighed alle døgnets 24 timer under behandlingsperioden, som varer 3-5 måneder. Mere end 250 terapeuter i Nord Amerika og Europa er certificerede MST-terapeuter.

Formålet med forskningsoversigten er at evaluere effekterne af MST på adfærdsmæssige dimensioner herunder kriminalitet, alkohol og stofmisbrug, skolefravær, på psykologiske dimensioner herunder psykiske symptomer og stress hos forældre, og på familiemæssige dimensioner som fx behandling uden for hjemmet og kvaliteten af børn-forældre-forholdet.

Figur 5.3 Et eksempel på et Campbell review

Faser:	**Eksempel: Multisystemisk terapi (MST) til behandling af sociale, følelses- og adfærdsmæssige problemer for unge i aldersgruppen mellem 10 og 17 år**
Problemstilling:	Hvilke er effekterne af multisystemisk terapi?
Søgning:	Der blev fundet 266 titler og abstracts, hvoraf 95 blev vurderet som relevante. På basis af læsning af de 95 rapporter blev 35 primærstudier identificeret.
Kritisk vurdering:	14 primærstudier blev ekskluderet, de fleste fordi de ikke var baseret på RCT, enkelte fordi de fokuserede på andre deltagertyper og en enkelt fordi den ikke præsenterede tilstrækkelige data til at det kunne danne basis for statistisk analyse. 13 primærstudier var endnu ikke færdiggjort og blev derfor lagt til side for efterfølgende vurdering. Tilbage var 8 primærstudier, der mødte inklusionskriterierne. Data blev uddraget vedrørende design, deltagere samt resultatmål. Statistisk analyse blev gennemført for studier, der inkluderede samme resultatmål.
Konklusion:	Der er ingen troværdig evidens for, at MST er et bedre virkemiddel end alternativerne på de fleste resultatmål. Der er heller ingen evidens for, at MST har skadelige virkninger.

Bem.: (Littel JH, Popa M & Forsythe B., 2005).

Som figuren viser, er praksis i fremgangsmåden i MST-eksemplet overensstemmende med praksis i Cochrane reviewet vedrørende antibiotisk forebyggelse. Der er klare paralleller i kriterierne for inklusion og eksklusion af primærstudier og de kriterier, studierne er vurderet på, og måden, hvorpå resultaterne er syntetiseret. RCT og metaanalyse er nøglebegreberne. MST-oversigten er da også registreret i både Campbells og Cochranes databaser.

I MST-oversigten er der gennemført ganske grundig kvalitetsvurdering af primæsstudier (Littel 2005). Studierne er ikke kun vurderet med hensyn til skjult tildeling til hhv kontrol- og indsatsgruppe, men også deres

generelle kvalitet er vurderet. Her er intention-to-treat analyser vurderet som værende af højere kvalitet end analyser af udbytte (outcome) for dem, der har gennemført programmet (TOT-analyse). TOT-analyse betragtes som tenderende til at overvurdere programmets resultater, idet den ekskluderer dropouts og nægtere, som generelt vil have mere negativt udbytte end dem, der gennemfører et program.

MST-forskningsoversigten konkluderer, at der ikke er evidens for, at MST er bedre end alternative behandlinger. Men oversigten giver også anbefalinger om, hvordan primære studier kan forbedres i fremtiden. Der argumenteres for, at primærstudier fremover bør baseres på RCT, som anvender blind randomisering, hvor det er muligt, og de bør udformes, så der kan gennemføres en intention-to-treat analyse.

Bag denne tekniske metodediskussion ligger en diskussion om uafhængighed. De fleste primærstudier af MST er blevet foretaget af dem, der har udviklet MST-programmet. Dette kan have ført til, at man har fokuseret på positive outcomes, fordi man har søgt (ubevidst) bekræftelse på, at programmet virkede. MST-oversigten er da også af programudviklere blevet kritiseret for at fejlfortolke og give en forkert fremstilling af en række af de primære MST-studier (se Henggeler et al. 2006, og svaret fra Littel 2006).

Der er imidlertid i Campbells oversigter eksempler på andre typer af praksis ved forskningsoversigter. Eksemplet i figur 5.5 drejer sig om aktivering af ledige. Oversigtens spørgsmål er, hvorvidt der er en trusselseffekt af aktiveringsforanstaltninger (som er tvungne). Eller med andre ord, finder de ledige, som skal aktiveres selv arbejde for at undgå aktiveringsforanstaltninger? I dette eksempel bygger protokollen ikke kun på primærstudier med RCT, men også studier, der er udformet som kvasieksperimenter, naturlige eksperimenter, og som anvender økonometriske analyser af registerdata (Bjørn et al. 2004a). Hvis man kun havde inkluderet RCT-studier, ville alle studier gennemført i de nordiske lande være blevet udelukket (i eksemplet 8 ud af 13 studier). Og det på trods af, at de nordiske lande regnes for foregangslande med hensyn til aktiv beskæftigelsespolitik.

Figur 5.4 Et eksempel på et Campbell review

Faser:	Eksempel: Beskæftigelseseffekt forårsaget af truslen om aktivering
Problemstilling:	Kan der dokumenteres en trusselseffekt i forbindelse med obligatorisk deltagelse i indsatser inden for aktiv arbejdsmarkedspolitik?
Søgning:	Der blev fundet 13 undersøgelser, der potentielt kunne bidrage med evidens.
Kritisk vurdering:	Alle undersøgelser blev inkluderet både 3 amerikanske undersøgelser baseret på RCT og en række undersøgelser fra andre lande baseret på andre typer af kontrolgrupper og forskellige former for komparative design. Data blev uddraget vedrørende design, deltagere, type af intervention samt resultatmål. På grund af heterogenitet studierne imellem, blev der ikke gennemført statistisk analyse, men i stedet en deskriptiv komparation (»narrative findings«).
Konklusion:	Hovedparten af de gennemførte studier finder evidens for en trusselseffekt.

Bem.: Et eksempel på et Campbell review (Bjørn N.; Geerdsen L. & Jensen P. 2004b).

I forskningsoversigten vedrørende trusselseffekt blev den metodologiske kvalitet af RCT-studier og andre typer af studier vurderet med afsæt i hvert sit sær af kriterier. RCT-studierne blev vurderet i forhold til randomisering, uafhængighed og frafald i forhold til den oprindelige stikprøve. De øvrige forskningsdesign blev vurderet på grundlag af tre kriterier: antagelserne om identifikationen af trusselseffekten, korrektion for uobserveret heterogenitet og korrektion for ikke tilfældig udvælgelse til programmet. Ifølge protokollen var planen at foretage metaregressionsanalyse ved syntese af primærstudiernes resultater (Bjørn et al. 2004a). Imidlertid argumenterer forfatterne i oversigten for at anvende narrativ syntese i stedet på grund af de primære studiers forskelligheder.

Eksemplet fra beskæftigelsespolitikken illustrerer et grundlæggende problem, som hænger sammen med forskelle i forskningstraditioner i hhv. USA og Europa. Når man bevæger sig uden for sundhedsområdet til socialområdet og uddannelsesområdet, er der relativt få RCT-studier i Europa. Hvis man derfor kun eller overvejende medtager RCT-baserede studier i forskningsoversigterne, som det er formuleret af både Cochrane og Campbell, så får man konklusioner, som næsten udelukkende er baseret på amerikanske erfaringer. Dette kan indebære, at man ikke får indhøstet erfaringer fra Europa, inkl. Norden. Og det indebærer, at man må forholde sig til spørgsmålet om, hvorvidt erfaringer fra andre kulturelle og social og politisk-administrative kontekster kan overføres.

Beskæftigelseseksemplet viste også, at Campbell Collaboration tillader, at reviewerne udvikler en praksis, der ikke kun er baseret på inklusion

af RCT-baserede primærstudier. Man accepterer også inklusion af kvasieksperimenter, naturlige eksperimenter og registeranalyser. Det er imidlertid værd at bemærke, at forskningsoversigten blev sendt til godkendelse december 2004, og at den endnu ikke (august 2007) er blevet godkendt. Det kan der naturligvis være mange grunde til, men en kunne formodes at være diskussionen om narrativ analyse frem for metaanalyse.

The Social Care Institute for Excellence (SCIE) bruger som nævnt betegnelsen »knowledge reviews« for deres vidensoversigter. SCIE's reviews trækker typisk på flere vidensbaser. Et review om, hvordan invalide og uarbejdsdygtige forældre bedst støttes, inkluderer fx både et omfattende litteraturreview, der i øvrigt suppleres med møder med grupper af forældre, der er underrepræsenteret i litteraturen, fx forældre med HIV/AIDS, og en survey udsendt til lokale myndigheder med formålet at afdække eksempler på god praksis (Morris & Wates 2006). Et andet review vedrørende individualiseret, brugerorienteret ældreservice inkluderer ligeledes både en forskningsoversigt og en survey til personer og organisationer involveret i sådanne indsatser samt seks casestudies af udvalgte lokale myndigheder (Glendinning m.fl. 2007). Forskningsoversigten begrænsedes eksplicit til nyere publikationer i England med det argument, at der ikke ville kunne generaliseres på tværs af lande og ej heller over tid på grund af hurtigt skiftende politiske såvel som praksiskontekster.

Som det fremgår, er praksis ved udarbejdelsen af forskningsoversigter mere forskelligartet på socialområdet end på sundhedsområdet. Både eksemplet med beskæftigelsespolitikken og eksemplerne fra SCIE's praksis illustrerer, at mens ideen om reviewpraksis er blevet spredt, er den blevet tilpasset den institutionelle kontekst og de metodologiske traditioner.

Eksempler fra uddannelsesområdet

Inden for uddannelsesområdet er det engelske EPPI-center ganske produktivt. Via EPPI's hjemmeside gives adgang til 56 forskningsoversigter på uddannelsesområdet (april 2007), hvor Campbell lister 18 projekter på uddannelsesområdet, hvoraf kun fem er offentliggjorte forskningsoversigter. EPPI udarbejder forskellige typer af forskningsoversigter. De anvendte syntesemetoder omfatter metaanalyse, narrative og begrebsmæssige synteser.

Figur 5.6 præsenterer et eksempel på en EPPI-oversigt, som opsummerer proces og resultater. Eksemplet omhandler den såkaldte National Numeracy Strategy (NNS), som blev indført i grundskolen i England i 1999. NNS omfattede følgende elementer: en daglig matematiktime, en tredelt struktur på disse timer, vægt på interaktiv undervisning med hele klassen. Oversigtens spørgsmål var: Har den daglige matematiktime hjulpet eleverne til at udvikle fortrolighed med og kompetencer i elementær matematik?

Figur 5.6 Et eksempel på et EPPI-review

Faser:	**Eksempel: Daglig matematiktime i grundskolen I England**
Problemstilling:	Har indførelsen af en daglig matematik time i grundskolen hjulpet eleverne til at udvikle sikkerhed og kompetence i tidlig matematik?
Søgning:	736 publikationer af potentiel relevans blev identificeret. Af disse blev 671 ekskluderet på titel og abstract. Kriterier for in- og eksklusion var relateret til publikationstype og relevans ikke til design and metodologi.
Kritisk vurdering:	Af de tilbageblevne 65 publikationer viste 2 sig at være mundtligt præsenterede konferencepapers, som ikke var tilgængelige i komplet skriftlig version. Af de tilbageblevne 63 blev andre 43 ekskluderet, da de viste sig ikke at indeholde vurderinger af effekten af den daglige matematiktime. De 20 publikationer præsenterede resultater fra 18 primærstudier. Resultaterne fra alle disse blev inkluderet. Blandt de 18 primærstudier var forskellige design blandt andet surveys, interview, observationsstudier og resultater fra test af elevernes matematiske kompetence. I den kritiske vurdering af de 18 studies blev 11 karakteriseret som værende af høj kvalitet, mens 7 blev karakteriseret som »medium«. Ingen primærstudier blev karakteriseret som værende af lav kvalitet. Resultaterne of primærstudierne blev sammenfattet narrativt.
Konklusion:	Reviewet konkluderer, at den daglige matematiktime er blevet godt modtaget og bredt implementeret, samt at der er nogen evidens for, at den har forøget elevernes sikkerhed og kompetence i tidlig matematik. De opnåede resultater kan dog afspejle en forøget overensstemmelse mellem det, der undervises i og det, der testes, snarere end forøgelse af elevernes forståelse af matematik.

Bem.: (Kyriacou and Goulding et al. 2004).

Som det fremgår af figuren, omhandler denne forskningsoversigt et nationalt engelsk policy-initiativ. Kriterier for valg af primærstudier går mere på relevans og mindre på forskningsdesign og metode. Forskningsoversigten omfatter primærstudier med forskellig metodologi, fx survey, interview, observationer og test. Resultaterne er sammenfattet ved narrativ syntese.

Det amerikanske What Works Clearinghouse (WWC), der blev etableret i 2002 af Department of Education's Institute of Education Sciences,

har udarbejdet syv forskningsoversigter. Disse omfatter tidlig læsning, førskoleundervisning, forebyggelse af dropout, elementær matematikundervisning, engelsk, personlighedsudvikling og matematikpensum på mellemniveau. WWC har en standardiseret fremgangsmåde for at udarbejde forskningsoversigter, som også omfatter kriterier for at vurdere kvaliteten af primærstudier og rangordne effektivitet af indsatser. Kvaliteten af primærstudier vurderes og rangordnes i tre kategorier (se figur 5.7).

Figur 5.7 WWC's kriterier for kvalitetsvurdering af primærstudier

Category	Definition
Meets evidence standards (MES)	RCT that do not have problems with randomisation, attrition or disruption, and regression discontinuity design that do not have problems with attrition or disruption.
Meets evidence standards with reservations (MESR)	Strong quasi-experimental studies that have comparison groups and meet other WWC evidence standards, as well as RCT with randomisation, attrition or disruption problems, and regression discontinuity design with attrition or disruption problems.
Does not meet evidence standards (DNMES)	All the rest.

WWC tilslutter sig evidenshierarkiet og kun primærstudier, som vurderes til MES eller MESR, inkluderes i forskningsoversigter. Indsatsers effektivitet vurderes i fem kategorier: positive, potentielt positive, blandet, ingen mærkbar effekt, potentielt negativ eller negativ. Ved rangordenen tages der hensyn til fire faktorer: forskningsdesignets kvalitet, fundenes statistiske signifikans, forskelle i størrelse i hhv forsøgs- og kontrolgruppen, og resultaternes konsistens på tværs af studier.

WWC har udviklet et såkaldt forbedringsindeks, som viser forskellen i procentrangordenen mellem den gennemsnitlige elev i indsatsen og sammenligningsgruppen. Indekset kan variere i værdi mellem 50 and +50. Positive værdier betyder gunstige resultater.

Figur 5.8 viser et eksempel på en forskningsoversigt vedrørende elementær matematikundervisning baseret på tre forskellige lærebogssystemer. Spørgsmålet var: Hvad er effekterne af de forskellige typer af matematikundervisning på elevernes præstationer? Hver type er rapporteret for sig.

Figur 5.8 Et eksempel på et WWC-review

Elementær matematikundervisning baseret på: **Faser:**	**»Everyday mathematics«**	**»Saxon elementary school math«**	**»Scott Foresman-Addison Wesley mathematics«**
Søgning	61 primærstudier blev identificeret	7 primærstudier blev identificeret	4 primærstudier blev identificeret
Kritisk vurdering	Antal studier scoret: – MES: 0 – MESR: 4	Antal studier scoret: – MES: 0 – MESR: 1	Antal studier scoret: – MES: 1 – MESR: 0
Konklusion	Forbedringsindeks: +12 Potentielt positive effekter på matematik præstationer	Forbedringsindeks: +7 Ingen mærkbare effekter på matematik præstationer	Forbedringsindeks: -2 Ingen mærkbare effekter på matematik præstationer

Som figuren viser, er der i eksemplet et meget varierende antal primærstudier vedrørende de tre lærebogssystemer. Fælles er det imidlertid at kun ganske få primærstudier bliver vurderet som værende af en kvalitet, der er acceptabel. Konklusionen af den statistiske syntese af de inkluderede primærstudiers resultater er, at der kun er potentielt positiv evidens for effekter i den ene type af undervisning. Det skal bemærkes, at WWC alene søger og inkluderer amerikanske primærstudier. Dette kan synes overraskende, al den grund at WWC er tæt knyttet til Campbell Collaboration, idet en af initiativtagerne til Campbell, Robert Boruch, også er chefundersøger i WWC. WWC's nationale profil synes at have sammenhæng med, at man tror, at amerikanske lærere kun vil fatte lid til amerikanske undersøgelser.

Figur 5.9 viser, hvordan WWC har kvalitetsvurderet alle de primærstudier, der er fundet via litteratursøgning i udarbejdelsen af de syv forskningsoversigter, der indtil dato er udarbejdet.

Tabel 5.1 Kvalitetsvurdering af primærstudier i WWC-reviews

Review: / Kategori:	Tidlig læsning	Førskole undervisning	Forebyggelse af dropout	Elementær matematikundervisning	Engelsk	Personlighedsudvikling	Matematik på mellemniveau	Total
MES	4	8	3	1	4	7	3	30
MESR	0	2	1	5	6	11	11	36
DNMES	1	9	10	66	4	37	62	189
Total	5	19	14	72	14	55	76	255

Bem.: (baseret på WWC's webside 24.11.06)

Figuren viser, at kun 12% af alle studier (30 ud af i alt 255 studier) vurderes at opfylde evidensstandarderne, 14% opfylder standarderne med reservation, og i alt 74% opfylder ikke standarderne. Der er stor variation på tværs af temaer for forskningsoversigter. Alt i alt sker der en betydelig selektion. En stor andel af foreliggende primærstudier dømmes bort.

5.2 Snittet for inklusion

I eksemplerne fra både Cochrane (sundhedsområdet), Campbell (socialområdet) og WWC (uddannelsesområdet) blev der i nogle forskningsoversigter kun medtaget RCT-studier, mens der i andre forskningsoversigt også blev medtaget ikke RCT-studier, dog kun studier placeret højt i evidenshierarkiet. I eksemplet fra EPPI blev inkluderet primærstudier med forskellige design ud fra relevansovervejelser og i mindre grad ud fra en rangorden af primærstudiers design. Det samme gjaldt eksemplerne fra SCIE, men her blev resultater fra allerede foreliggende undersøgelser endvidere kombineret med ny dataindsamling. Eksemplerne peger altså på en betydelig variation i, hvilke design af primærstudier man accepterer for inklusion i en forskningsoversigt.

Talmæssige opgørelser fra forskningsoversigter fra hhv. Cochrane, Campbell og EPPI bekræfter denne variation og viser, at de evidensproducerende organisationer anvender forskellige kriterier. En opgørelse af alle Cochrane forskningsoversigter publiceret i året 2005 (i alt 497) viser, at 69% af alle forskningsoversigter inkluderer alene RCT-designede primærstudier. En lignende opgørelse af Campbell forskningsoversigter (alle

forskningsoversigter publiceret frem til marts 2007, i alt 20) viser, at 50% af disse inkluderer alene RCT-designede primærstudier. Heroverfor har EPPI (omfattende alle forskningsoversigter frem til marts 2007, i alt 74) ingen review, som kun medtager RCT.

5.3 Kvalitetsvurdering af primærstudier

Som det fremgår af eksemplerne, gennemføres kvalitetsvurdering af primærstudierne ud fra forskellige vurderingsskalaer alt efter, hvilken evidensproducerende organisation der er tale om. Men selv inden for samme evidensproducerende organisation er der eksempler på, at der anvendes forskellige skalaer. Det gælder Cochrane-eksemplet med psykologisk debriefing for at forebygge posttraumatisk stress. Hvor der i den selv samme oversigt foretages vurdering efter flere skaler, for efterfølgende at vælge en skala, som var tilpasset (kontekstualiseret) til forskningsoversigtens emne. Vurderingskriterierne blev dog holdt inden for en RCT-logik.

Det har ikke været muligt inden for dette studies rammer at foretage en mere systematisk analyse af, i hvilket omfang hvilke vurderingsskalaer er blevet benyttet ved udarbejdelsen af forskningsoversigter, og hvordan det er blevet gjort i praksis. Litteraturen om vurderingskriterier for forskning og evalueringsstudier er imidlertid ganske omfattende, men vi er ikke vidende om empiriske studier af, hvordan kvalitetsvurderingen i praksis finder sted. I relation til forskningsoversigter gennemgår Petticrew og Roberts (2006) i kap. 5 forskellige skalaer og tjeklister, der anvendes til at vurdere primærstudier med forskellige design (RCT, forløbsundersøgelser, casestudier mv.). Forfatterne beskriver to tilgange til kvalitetsvurdering af forskning. Denne ene går tilbage til Campbell og kollegers arbejde med at identificere trusler mod gyldighed i kvasieksperimentelle studier (Campbell og Stanley 1966). Den anden tilgang blev oprindeligt udviklet af Thomas Chalmers og kolleger til vurdering af RCT-studier.

Endelig vil vi nævne en tredje tilgang til vurdering af primærstudier, som stammer fra litteraturen om, hvordan man kan sikre og forbedre kvaliteten af evalueringsstudier (Schwartz og Mayne 2005). Denne tilgang rummer flere aspekter af kvalitetssikring. Her omtales ud over diverse tjeklister og håndbøger til vurdering af færdige enkeltstudier også vurdering i løbet

af gennemførelsen af et evalueringsstudie samt vurdering af systemer til kvalitetssikring (audits, certificering etc.).

I den her nævnte litteratur findes også en diskussion af fordele og ulemper ved systematisk vurdering af forskning og evalueringsstudier ved tjeklister og andre standardiserede redskaber. På den ene side fremhæves, at kvalitetsvurderinger uden tjeklister fører til mindre kritisk vurdering, end hvis man anvendte tjeklister (Petticrew og Roberts 2006, s. 154). På den anden side advares mod at lade standardiserede tjeklister styre vurderingen for meget. Der bør være plads for det faglige skøn og til at foretage helhedsvurderinger, som går ud over vedtagne metodestandarder. Fx bør der skelnes mellem kvaliteten ved rapporteringen (som kan mangle metodeoplysninger) og selve studiets kvalitet (Petticrew og Roberts 2006, s. 126f).

5.4 Syntetisering af resultater

Som nævnt i afsnit 4.3 betragter de organisationer, der prioriterer primærstudier med RCT-design metaanalyse som den ideelle syntetiseringsmetode. Også på dette område er der overensstemmelse mellem politik i form af retningslinjer og praksis. En opgørelse af alle Cochrane-forskningsoversigter publiceret i året 2005 (i alt 497) viser, at 68% af alle forskningsoversigter som udgangspunkt benytter metaanalyse, for Campbell (alle forskningsoversigter publiceret frem til marts 2007, i alt 20) er tallet 60%, mens det for EPPI (omfattende alle forskningsoversigter frem til marts 2007, i alt 74) kun er 9%. For både Campbell og EPPI benyttes metaanalyse herudover i nogle forskningsoversigter i kombination med andre syntetiseringsformer, typisk narrativ syntetisering.

5.5 Det pragmatiske perspektiv: Hvilken evidens er tilgængelig og hvilken accepteres?

Som det er fremgået, er der en vis variation i reviewpraksis mellem de evidensproducerende organisationer. Nogle organisationer følger evidenshierarkiets rationale, mens andre arbejder ud fra en bredere evidensbase. For alle organisationer gælder det, at der er god overensstemmelse

mellem de retningslinjer og politikker, de har udarbejdet, og den praksis, de anlægger. Det er dog tydeligt, at der ofte også er en række mere pragmatiske forhold, der har betydning for, hvordan konkrete reviews udarbejdes. Det drejer sig frem for alt om, hvilken type evidens der er tilgængelig i en konkret situation og, i nogle evidensorganisationer, hvilken type af evidens der formodes at blive accepteret hos brugerne af resultaterne af systematiske forskningsoversigter.

Som tidligere nævnt foretrækker en del af evidensbevægelsens organisationer i forbindelse med studier af interventionseffekter primærstudier, der er designet som RCT. Såfremt sådanne ikke er tilgængelige, eller kun er tilgængelig i begrænset omfang, er der overordnet set to strategier – enten kan man konkludere, at der ikke findes relevant evidens om den undersøgte intervention, eller også kan man acceptere »ringere« forskningsdesign, dvs. design længere nede i evidenshierarkiet som fx kvasieksperimenter og forløbsundersøgelser med statistisk kontrol for dog at frembringe en analyse af bedst *tilgængelige* evidens.

Det skal dog påpeges, at det ikke i alle tilfælde vil være muligt at tage højde for, hvilken evidens der reelt er tilgængelig. Dette gælder fx i forbindelse med Cochrane- og Campbellorganisationernes systematiske reviews, hvor de væsentligste valg omkring inklusion af studietyper foretages i en såkaldt reviewprotokol forud for litteratursøgningen. Inklusionskriterierne fastlægges således, inden evidensbasen på området er kendt. Oftest vil de, der udarbejder protokollen, dog have forhåndskendskab til området, så de på forhånd har en vis idé om, hvor stor evidensbasen på området er, og derfor kan indarbejde dette i deres overvejelser.

Reviewpraksis er i øvrigt en metodologi i udvikling. Der er for det første øget fokus på, hvad der kendetegner god kvalitet i reviewpraksis. Denne diskussion vedrører blandt andet vigtigheden af gennemsigtighed i de valg, der foretages i afgrænsningen af reviews (se Moher m.fl. 2007 og Schlosser 2007). Herudover er der en fortsat diskussion af inklusionskriterier. Inklusionskriterier diskuteres løbende. I den nyligt gennemførte evaluering af Nordisk Campbell Center (NC2) anbefales det for eksempel, at NC2 bør arbejde for, at den internationale Campbell organisation bliver mere åben over for andre kvantitative undersøgelsesdesign end RCT (Fisker m.fl. 2007). Tankegangen er, at evidensbasen som udgangspunkt bør

inkludere alle kvantitiative undersøgelssdesign inklusive økonometriske metoder, samt at kvaliteten af de forskellige design bør vurderes på hver deres sæt af kriterier. Konsekvensen af at følge disse anbefalinger vil således til dels være, at tankegangen om evidenshierarkiet forlades, idet evidensbasen dog alene udvides med kvantitative design.

Ud over begrænsningerne i de tilgængelige primærstudiers metodologi er der den begrænsning, der fremkommer ved, at den primære målgruppe for forskningsoversigten anses for kun at ville accepterer bestemte former for viden. Det ses tydeligt i relation til WWC, der på trods af slægtskabet til Campbell med afsæt i, hvad brugerne forventes at acceptere, fravælger ambitionen om at producere universel viden. Det ses ligeledes på sundhedsområdet, hvor det fremherskende videnskabsteoretiske paradigme (af nypositivistisk karakter) tilsyneladende begrunder (måske ubevidst for den enkelte), hvor snittet lægges. I nogle vejledninger synes et bestemt paradigme at dominere, når det fremgår, at der inden for evidensbevægelsen på sundhedsområdet er enighed om, at RCT er det forskningsdesign, der skaber mindst bias, men til gengæld er det også klart, at desto strengere kvalitetskrav der stilles til de inkluderede studier, desto sværere er det at finde tilstrækkelig med relevante studier til at foretage en reel vurdering af interventionen. Cochrane Collaboration stiller i deres håndbog problemet således op:

How far is it possible to achieve a higher level of relevance by including evidence other than that derived from RCT without violating the central principle: minimising bias? (Cochrane-håndbogen citeret i Ogilvie 2005).

Man erkender altså i Cochrane-håndbogen, at der er et trade-off, men giver ikke noget direkte svar på, hvad man skal vælge som cut-off point (Ogilvie 2005).

6 Argumenter for og imod RCT som den gyldne standard

Som det er fremgået af det foregående, er der evidensproducerende organisationer, der altovervejende prioriterer RCT som det design, der producerer valid viden, og andre evidensproducerende organisationer, der arbejder ud fra en en bredere evidensbase. I USA er der i de senere år oprettet organisationer, der går stærkt ind for RCT, som fx Coalition for Evidence-Based Policy. I Europæisk sammenhæng er der især på sundhedsområdet via Cochrane samarbejdet og på velfærdsområdet, uddannelsesområdet og det kriminologiske område via Campbell samarbejdet blevet argumenteret for anvendelsen af RCT. Debatten for og imod RCT er ofte ganske ophedet. Kritikere er især at finde inden for det sociale område (se i dansk sammenhæng diskussionen i Social kritik, nr. 102, december 2005) og på uddannelsesområdet (se i dansk sammenhæng Moos m.fl., 2005 samt diskussionen i Unge Pædagoger, nr. 3, juli 2006). På uddannelsesområdet har OECD holdt en række konferencer. Også her er det klart fremgået, at opfattelserne er forskellige (OECD, 2004). I det følgende sammenfatter vi argumenterne for og imod RCT som det principielt bedste forskningsdesign, og vi ser på, hvordan evidensbevægelsen har givet svar på den rejste kritik.

6.1 Argumenter for RCT

Som tidligere nævnt er det bærende princip i RCT den tilfældige tildeling af individer (eller andre undersøgelsesenheder) til henholdsvis en gruppe, der udsættes for indsatsen (eksperiment- eller interventionsgruppen), og

en gruppe, der ikke får indsatsen (kontrolgruppen), samt registrering af effekter i begge grupper både før og efter, indsatsen er gennemført. Dette princip sikrer, at alle andre mulige årsagsforhold end den aktuelle indsats, som kunne have en effekt, er ligeligt fordelt (inden for statistiske rammer) i de to grupper. Dermed er den målte nettoeffekt netop effekten af den aktuelle indsats, idet alle andre faktorer er »holdt konstante« gennem den tilfældige, ligelige fordeling af individer i de to grupper. RCT-designet er på denne baggrund det design, der bedst løser isoleringsproblemet (Vedung 1998: 137). RCT tager i modsætning til fx økonometriske forløbsstudier via randomiseringen også hensyn til ikke erkendte (ikke observerede) mulige influerende faktorer.

Det fremhæves endvidere, at sammenligning af indsatsers effekter undersøgt ved både RCT og andre forskningsdesign, som ikke er RCT, viser, at RCT i højere grad påviser lavere eller ingen effekter, end tilfældet er for ikke-RCT-design. RCT er altså en »strengere« test. Det hævdes derfor, at RCT-design har mindre »bias«, dvs. er mindre fejlbehæftede, end ikke-RCT design, at årsag-virknings-afdækningen altså er mere »korrekt«. RCT-designet karakteriseres på denne baggrund som et design, der sikrer høj intern validitet.

Det understreges samtidig, at der stilles krav til det stærke RCT-design, jf. figur 6.1.

Figur 6.1 Karakteristika ved det stærke RCT-design

Kendetegn ved det stærke RCT-design
• Tilfældig tildeling af individer (enheder) til indsats- eller kontrolgruppen • Blinding for at sikre, at ingen parter kender til, hvem der er i hhv. indsats- og kontrolgruppen • Sammenligning af grupperne, før indsatsen påbegyndes (baseline målinger) • Fuldstændig opfølgning af alle, der får, og alle, der nægter at modtage indsatsen • Objektiv og neutral (unbiased) vurdering af outcome • Analyse baseret på oprindelig gruppefordeling • Vurdering af sandsynligheden for, at de målte resultater alene skyldes tilfældigheder • Vurdering af den statistiske styrke i undersøgelsen

Bem.: (Davies, Nutley & Tilley, 2004: 260).

Fortalere for dette synspunkt er imidlertid også opmærksom på de grænser, RCT har for i praksis at kunne leve op til de ideale principper, fx nævner Farrington (2003: 53f), at man også bør undersøge de kausale

kæder mellem intervention og outcome og vurdere gyldigheden af teori herom.

6.2 Argumenter imod RCT

Der er fremsat mange argumenter mod og kritik af RCT som det principielt bedste forskningsdesign. Vi har kategoriseret argumenterne i nogle hovedkategorier, som strukturerer dette delafsnit, hvor vi giver en sammenfatning af argumentationen. Det er tilstræbt at præsentere eksempler på argumenterne, de er ikke dækkende eller repræsentative for debatten, som foregår både i sektororienterede (fx uddannelse, kriminalforsorg), i disciplinorienterede (fx lægevidenskab, pædagogik og sociologi) og i tværgående tidsskrifter og fora (fx evalueringstidsskrifter og -konferencer).

6.2.1 Tekniske problemer

For det første kan der peges på argumenter, der relaterer sig til en række tekniske problemer. Disse relaterer sig først og fremmest til følgende metodologiske begrænsninger knyttet til randomisering og sikkerhed i måling (se endvidere Davies, Nutley & Tilley, 2004: 262; Feinstein & Horwitz, 1997; Launsøe og Gannik, 2000; Mullen et al., 2004; Sauerland 1999):

- Manglende blinding. I relation til mange interventioner er det vanskeligt eller umuligt at tilrettelægge blindede forsøg. Dette gælder på dele af det medicinske område (fx kirurgien) og er udbredt på en række andre områder, fx velfærds- og uddannelsesområdet. Fravær af blinding introducerer risiko for performancebias, jf. afsnit 4.2.
- Individuelle præferencer. Patienters/klienters individuelle præferencer for bestemte behandlinger kan reducere det antal, der ønsker at deltage i et forsøg, og dermed vanskeliggøre randomisering. Individuelle præferencer kan ligeledes være årsag til frafald undervejs, hvilket kan reducere validiteten. Randomisering viser sig således ofte at være ufuldstændig i praksis.
- Varierende leverance. I mange situationer påvirkes leveringen af en indsats af behandlerens faglighed og faglige skøn. Variationer i faglige

normer kan vanskeliggøre standardisering. Manglende standardisering af indsats introducerer usikkerhed i forhold til, hvad der egentlig evalueres.

- Brugertilpassede interventioner. Nogle typer af indsatser kræver tilpasning til den individuelle bruger for at sikre effekt. Brugertilpasning vanskeliggør standardisering og introducerer usikkerhed i forhold til, hvad der egentlig evalueres.
- Interaktion. Interaktion mellem behandler og klient kan påvirke visse typer af interventioners effekt.
- Svag efterlevelse (compliance). Tager patienterne/klienterne ikke behandlingen alvorligt (tager de fx ikke den ordinerede medicin, eller følger de ikke træningsprogrammet), undermineres forsøget. Det samme er tilfældet, hvis personer i kontrolgruppen »snyder« og alligevel modtager interventionen måske fra et andet program.
- Ressourcer. Randomisering kræver en del organisering og er derfor et relativt tidskrævende og dyrt undersøgelsesdesign.
- Måling. Der måles kun i en given tidsperiode. Det er usikkert, om effekterne er varige. Indsamles resultaterne via patienternes/klienternes selvrapportering, er der risiko for bias.

Ud over ovenstående kan nævnes den problemstilling, der knytter sig til, hvornår i en udviklingsfase af en intervention RCT-undersøgelsesdesignet med fordel kan anvendes. Tilrettelægges et RCT tidligt i en interventions udviklingsfase, inden den så at sige er kommet sig over sine børnesygdomme, løbes en risiko for, at indsatser med potentiale forkastes som værende uden effekt.

6.2.2 Etiske problemer

Kritikere af RCT har også rejst en række etiske problemstillinger. For det første argumenterer de for, at man ved blinding bedrager forsøgsdeltagerne. Godt nok kan »bedraget« imødegås ved deltagernes accept af at deltage i forsøget (det såkaldte informerede samtykke), men spørgsmålet er, om informeret samtykke er tilstrækkeligt til at løse de etiske problemer især i tilfælde, hvor der er tale om indgribende interventioner.

Kritikerne argumenterer endvidere for, at der er et etisk problem knyttet til, at individet ved at indgå i et RCT fratages en del egenkontrol. Fx er det ofte en betingelse i medicinske forsøg, at patienter ikke må modtage alternative behandlinger under forsøget. Herudover er det ifølge kritikerne svært at retfærdiggøre, at man ikke vil give kontrolgruppen den samme behandling som forsøgsgruppen (som antages at være bedre end den sædvanlige behandling, selv om man jo ikke ved det ved forsøgets start). Modsætningsvis udsætter man forsøgsgruppen for risici, som er ukendte i udgangspunktet.

6.2.3 Smal evidens

Et tredje kritikpunkt er, at RCT-design resulterer i produktion af »smal« evidens. Via RCT-design skabes alene viden om, hvad der virker, ikke om hvorfor det virker eller ikke virker, og ikke om, hvordan de, der modtager indsatsen, oplever denne. Herudover er fokus oftest alene på effektivitet og ikke på omkostningseffektivitet. Der skabes med andre ord viden om, hvilke interventioner der virker henholdsvis ikke virker, men normalt ikke om deres relative omkostningseffektivitet. Værdier integreres ikke i analysen, hverken patient/klientværdier eller samfundsmæssige værdier. Set fra et praksisprofessionelt synspunkt og et beslutningstagersynspunkt kan dette synes snævert.

6.2.4 Kompleksitet, kontekst og dynamik

Nogle af de kritikpunkter, der fremstilles som tekniske problemer, reflekterer mere principielle problemstillinger knyttet til indsatser karakteriseret ved kompleksitet og dynamik. Hovedargumentet er her, at en del medicinske interventioner og de fleste sociale og uddannelsesmæssige interventioner er sammensatte, ikke-standardiserede interventioner, at årsagsvirknings-kæderne er komplekse og dynamiske, og at de gennemføres under kontekstspecifikke vilkår.

At konteksten er specifik har som konsekvens, lyder argumentet, at interventionens effekt vil være forskellig i forskellige kontekster. Hvis et RCT-design gennemføres med en relativt stor population i forskellige kontekster, fanges dette ikke ind, da kontekstens betydning skjules i gennemsnitstallene. Hvis RCT således viser, at en intervention ikke har effekt,

kan det skjule, at interventionen har positiv effekt i nogle kontekster, og negativ effekt i andre.

At interventioner ofte er sammensatte og dynamiske, betyder, at RCT ikke er relevant, idet RCT forudsætter en veldefineret, ensartet og stabil intervention (Hammersley 2005). Hvis interventionen ikke er den samme (teknisk identificerbar og standardiseret) for alle i populationen, vil der være tale om forskellige »årsager« i RCT-opstillingen. Her kan man notere sig, at RCT har fået et kraftigt opsving efter midten af det 20. århunderede, hvor medicinalindustrien begyndte at markedsføre og teste standardiserede lægemiddelbehandlinger. Dertil kommer, at hvor interventionen vedrører meget heterogene (forskelligartede) populationer, introduceres yderligere en kompleksitet, som på sundhedsområdet viser sig ved, at patientpopulationen er forskelligartet, såkaldt prognostisk inhomogenitet. Det indebærer, at RCT igangsættes over for en population, hvor sygdomme kan udvikle sig forskelligt, fordi individerne har forskellige diagnoser.

Et andet argument, som knytter sig til individerne, er, at de sjældent er passive modtagere af indsatser. De handler selv og reagerer på uforudsete måder. Kritikerne karakteriserer i denne sammenhæng RCT for at bygge på en behavioristisk stimulus-respons-model, som ikke tager hensyn til de ressourcer, det enkelte individ besidder, og som kan trigges af interventionen. Kausalitetsforståelsen karakteriseres som forsimplet, idet kausaliteten alene tillægges eksterne forhold og ikke interne forhold knyttet til individet (Scocozza 2000). Disse kritikpunkter er forsøgt imødegået ved blinding, dvs. at individerne ikke er klar over, om de får en intervention. Blinding er imidlertid vanskelig ved de fleste interventioner uden for lægemiddelbehandlingsområdet.

At årsagskæderne er komplekse, skulle ikke i sig selv være et argument mod RCT, da det er nettoeffekten, man opgør. Men i sammenhæng med at årsagskæderne kan forandre sig (er dynamiske) og er ukendte (black box) i RCT-undersøgelser, vanskeliggør det forudsigelse og anvendelse af RCT-resultater. Inden for evalueringslitteraturen er kompleksitetsargumentet især blevet fremført med styrke af Pawson (2006).

6.2.5 Ekstern validitet

Kritikerne argumenterer for, at den eksterne validitet af resultaterne fra RCT-design er lav. Problemet er, at anvendelsen af resultater fra RCT vanskeliggøres af, at RCT ofte foregår på nogle betingelser, som er forskellig fra de betingelser, hvorunder resultaterne efterfølgende skal anvendes. Forskellighederne kan knytte sig til, at populationen er anderledes, eller til, at kulturen omkring indsatserne er anderledes, fx at RCT viser en gennemsnitseffekt, men at professionelle har en individuel orientering til brugerne.

På sundhedsområdet har det vist sig, at patientpopulationen i RCT sjældent er repræsentative for hele befolkningen, idet kriterier for in- og eksklusion betyder, at kun 10-20% af almindelige patienter i normal, klinisk praksis indgår i RCT-forsøg.

6.2.6 Kausalitetsforståelse og videnskabsteoretiske argumenter

Bag ved ovennævnte argumenter for og imod kan man skimte forskellige kausalitetsforståelser med rod i forskellige videnskabsparadigmer. Disse paradigmer bliver imidlertid sjældent tydeliggjorte i debatten omkring evidens, og det vil føre for vidt i denne sammenhæng at behandle paradigmedebatten nærmere. Men det skal kort nævnes, at argumentationen for RCT bygger på en empirisk-analytisk videnskabsopfattelse, hvor der søges efter lovmæssigheder (herunder stabile årsag-virkningsforhold), som antages at eksistere »objektivt«, dvs. adskilt fra det iagttagende subjekt (Radnitzky 1970, Habermas 1968).

I en klassisk videnskabsteoretisk-metodologisk diskurs (Hammer 2004, Hansen 2003, Launsø og Rieper 2005) fremhæves andre paradigmer, som fx det hermeneutiske/fortolkende og det kritiske paradigme. I evalueringslitteraturen har man på grundlag af »realist theory« (Bhaskar 1978) udviklet fremgangsmåder i systematiske forskningsoversigter, som bygger på en anden kausalitetsforståelse, som er blevet benævnt konfigurativ kausalitet (Pawson 2006). Hovedargumentet kan kort præsenteres som følger: Interventioner er teorier om, hvad der skal til for at forbedre forhold. De »rejser langt«, dvs. starter i hovedet på beslutningstagere, opbygges og virkeliggøres i samspil med interessenter (policy-design) og ændres under-

vejs. De individer (eller organisationer), som de er rettet imod, kan forholde sig aktivt påvirkende til interventionen. Interventionen vil få forskellige udfald (effekt) alt efter, hvilke målgrupper (og andre vilkår) den er underlagt. Der er således ikke en enkel, lineær dosis-respons-kausalitet mellem intervention og effekt, men en konfigurativ kausalitet, hvor »helheder« samvirker (what can you build with thousand of bricks? Lipsey 1997).

I det hele taget har debatten inden for evalueringsforskningen illustreret, at betydelige (videnskabs-) paradigmatiske forskelle gør sig gældende mellem fortalere for RCT og deres kritikere. En af de tydeligste kritikere i evalueringssammenhæng er som nævnt Pawson (2006), som også fremlægger alternativer til RCT-tænkningen både med hensyn til design af primærundersøgelser og udarbejdelse af systematiske forskningsoversigter.

6.3 Evidensbevægelsens svar på kritikken

Har evidensbevægelsen formuleret svar på den rejste kritik? Ja til dels.

Om de tekniske problemer kan man sige, at disse løbende er til debat. For det første forsøges de tacklet via diskussionen om forskellige typer af bias, og hvilke tiltag der kan sættes i værk for at imødegå disse. For det andet forsøges de tacklet ved at åbne op for inklusion også af andre typer af kvalitativt gode design fra den øverste ende af evidenshierarkiet. At anvendelsen af RCT-design skulle være særligt ressourcekrævende, modsiges herudover med afsæt bl.a. i et tidsperspektiv. På kort sigt er RCT-design måske dyrere, men på længere sigt er gevinsten stor, er argumentet. Fordi RCT-designet tilvejebringer mere gyldig viden, bliver der behov for færre undersøgelser totalt set (Cook 2002 og 2003). Kan der (delvis) anvendes allerede eksisterende data, bliver det eksperimentelle design i øvrigt endnu mere omkostningsefficient, selv om registeranalyser heller ikke er billige.

Kritikken vedrørende de etiske problemer har evidensbevægelsen formuleret et klart svar på. Argumentet er, at kritikkerne er galt afmarcheret. RCT-designet er etisk forsvarligt (Chalmers 2003). Faktisk fremhæver fortalerne for RCT, at det er uetisk ikke at foretage RCT med den begrundelse, at indsatser, der gennemføres uden RCT-baseret viden, i virkeligheden er et endnu større eksperiment, der blot ikke er kontrolleret. Man risikerer derfor at give mennesker en dårligere eller ligefrem skadelig indsats

uden viden fra RCT. Det er med andre ord uetisk at undlade at undersøge, om ens metoder virker. På det sociale område og på uddannelsesområdet er sagen i øvrigt ikke den, at man berøver nogle en virkningsfuld behandling. På disse områder vil kontrolgruppen typisk modtage den allerede eksisterende indsats, mens interventionsgruppen modtager en ny indsats, som man er usikker på effekten af, men har en formodning om er effektiv.

Kritikken relateret til kompleksitet og betydningen af kontekst karakteriseres af evidensbevægelsen som overdreven (Farrington 2003:65). Det afvises ikke, at kontekst kan have en betydning, men det understreges, at der ikke er nogen grund til at tage afsæt i, at kontekst altid har betydning. Spørgsmålet om, hvorvidt effekter af indsatser varierer med kontekst, bør derfor undersøges empirisk, men, lyder evidensbevægelsens argument, strategien må altid være først at vurdere den overordnede effekt af en indsats for eksempel via metaanalyse og derefter analysere moderatorers (herunder konteksters) påvirkning af denne.

På kritikken vedrørende kausalitetsforståelse og videnskabsteoretiske argumenter er der derimod ikke mange svar fra evidensbevægelsen. Denne kritik opfattes som ideologisk snarere end som metodologisk. Da den ikke opfattes som påvisende substantielle problemer, opleves den som ukvalificerede skældsord (Hede 2007). Dette opfattes ofte af kritikerne som en noget arrogant afvisning. Derved bliver debatten mere ophedet, end godt er.

6.4 Opsummering

Diskussionen de sidste år om styrker og svagheder ved RCT som det mest ideelle design til at afdække effekter af interventioner peger på en betydelig skepsis for og kritik af at gøre evidenshierarkiet med RCT på toppen til en universel retningslinje. En rapport til den amerikanske kongres fra eksperter afspejler ganske godt den debat i Europa, som vi har kendskab til. Nedenfor gengives et udpluk fra rapportens sammenfatning:

Views about the practical capabilities and limitations of RCT, compared to other evaluation designs, have sometimes been contentious. There is wide consensus that, under certain conditions, well-designed and implemented RCT provide the most valid estimate of an intervention's impact,

and can therefore provide useful information on whether, and the extent to which, an intervention causes favorable impacts for a large group of subjects, on average. However, RCT are also seen as difficult to design and implement well. There also appears to be less consensus about what proportion of evaluations that are intended to estimate impacts should be RCT and about the conditions under which RCT are appropriate. Many observers argue that other types of evaluations are necessary complements to RCT, or sometimes necessary substitutes for them, and can be used to establish causation, help bolster or undermine an RCT findings, or in some situations validly estimate impacts.

There is increasing consensus that a single study of any type is rarely sufficient to reliably support decision making. Many researchers have therefore embraced systematic reviews, which synthesize many similar or disparate studies. (Congress and Program Evaluation: An Overview of Randomized Controlled Trials (RCT) and Related Issues. March 7, 2006).
http://digital.library.unt.edu/govdocs/crs//data/2006/upl-meta-crs-9145/RL33301_2006Mar07.pdf?PHPSESSID=354150453e04b804d64d5e0e1be295b3

Det bemærkes, at et stridsspørgsmål er, under hvilke omstændigheder et veldesignet og gennemført RCT i det hele taget vil give gyldig og anvendbar viden. Og det bemærkes, at der er enighed om, at viden fra et enkelt studie sjældent er tilstrækkelig i praktisk og politisk sammenhæng, men at det kræver forskningsoversigter, som sammenfatter mange enkeltstudiers resultater.

7 Typologi over forskningsdesign

Et alternativ til at tage udgangspunkt i evidenshierarkiet, og et efter vores opfattelse mere konstruktivt startpunkt, er at tage udgangspunkt i en typologi. Evidenstypologier eller matricer angiver skematisk de enkelte undersøgelsesdesign udsagnskraft i relation til givne problemstillinger. Et eksempel på en sådan typologi er vist i figur 7.1 (opstillet på basis af Petticrew and Roberts 2003 og 2006: 60).

Figur 7.1 Et eksempel på en evidenstypologi

Design: / Research question:	Qualitative research	Survey	Case control studies	Cohort studies	RCT	Quasi-experimental studies
Effectiveness Does this work? Does doing this work better than doing that?				+	++	+
Process of service delivery How does it work?	++	+				
Salience Does it matter?	++	++				
Safety Will it do more good than harm?	+		+	+	++	+
Acceptability Will children/parents be willing to or want to take up the service offered?	++	+			+	+
Cost-effectiveness Is it worth buying this service?					++	
Appropriateness Is this the right service for these children?	++	++				
Satisfaction with the service Are users, providers and other stakeholders satisfied with the service?	++	++	+	+		

Bem.: Jo flere krydser jo større udsagnskraft har det pågældende design i forhold til den nævnte problemstilling.

Ifølge typologien har RCT således stor udsagnskraft i forhold til spørgsmål om effekter inklusive negative effekter (jf. begrebet »safety«). Herudover har de potentiale til at generere viden om omkostningseffektivitet, en dimension, der dog som tidligere nævnt kun sjældent indarbejdes i praksis. Ønskes der modsat sat fokus på, hvorfor en indsats virker eller ikke virker, og/eller på de involverede aktørers oplevelse af indsatsen, bør andre undersøgelsesdesign tages i brug.

Anvendelsen af evidenstypologier kan give inspiration til konstruktion af undersøgelsesdesign, der resulterer i mere helhedsorienterede perspektiver på givne indsatser. Lad os illustrere dette med et eksempel.

Et eksempel på kombination af design på det sociale område: En forskningsbaseret evaluering af rehabiliterings- og træningsindsatsen for børn med autisme, herunder evaluering af behandlingsmetoden ABA (Applied Behavior Analysis)

Rapporten omhandler resultaterne fra en evaluering af førskoletilbuddene til børn med autismespektrumforstyrrelser (ASF) i Danmark med særlig fokus på forsøget med ABA-metoden i Københavns Kommune (Høgsbro, 2007). I evalueringen indgår børnene fra fire typer af tilbud, som anvendes som kontrolgrupper i forhold til hinanden. Der belyses bl.a. følgende spørgsmål om både indsatserne og processerne i indsatserne og effekter af de forskellige former for behandling. Er der i praksis væsentlige forskelle mellem de forskellige modeller for indsatsen, eller er udvekslingen af delelementer i indsatserne så udbredt, at forskellene udviskes? Hvilke forskelle er der i effekten af de forskellige former for indsats? Påvirkes børnene, hvad angår IQ-præstation og autismespecifikke funktionsevnenedsættelser? Hvilke forskelle ses i forældrenes tilfredshed og belastning? Er indsatsens effekt knyttet til forældrenes forventninger til de enkelte modeller, måden modellen forklares på samt den løbende kontakt, dialog og samarbejdet mellem professionelle, forældre, andre børn, andet personale og pårørende? Hvilke økonomiske ressourcer lægger de enkelte modeller beslag på, og hvilke ressourcer frigøres som en effekt af den enkelte model? Frigøres der ressourcer hos forældrene, eller aflastes andre dele af de sociale, sundhedsmæssige og pædagogiske tilbud?

Evalueringen baseres på data fra såvel psykologiske test som kliniske iagttagelser, afrapporteret brugertilfredshed og observerbare ændringer i sociale konfliktniveauer, relationer og kompetencer. Der indgår således både kvalitative interview og spørgeskemaer til professionelle og forældre, samt observationer af børn og træningsforløb, ud over standardiserede test. Endvidere er der indsamlet økonomiske data fra kommunerne samt data vedrørende den kommunale visitationspraksis.

Evalueringens konklusion var, at ABA-metoden (som den dyreste behandlingsmetode) ikke var metoderne i de andre tilbud overlegen: »Evalueringen må på det foreliggende grundlag konkludere, at det enkeltintegrerede ABA-tilbud, som det er tilrettelagt i København, ikke har understøttet børnenes udvikling mere end det almindelige tilbud i Danmark.

Samtidig har det dog realiseret en støtte til familierne og en inddragelse af forældrene, som forældrene mangler i det almindelige tilbud. Og evalueringen efterlader et åbent spørgsmål om de forskellige pædagogiske princippers betydning hvad angår evne til social kontakt«.

Denne konklusion gav anledning til kritik fra en forældrekreds, som er tilhænger af ABA-metoden (ABAforum 2007).

For det andet kan anvendelsen af evidenstypologier inspirere til mere helhedsorienterede perspektiver på udarbejdelsen af forskningsoversigter. Evidensproducerende organisationer, der tager afsæt i evidenshierarkiet synes at specialisere sig på produktion af smal viden om effekter af interventioner. Spørgsmålet er imidlertid, om ikke brugere i politik og praksis kan drage mere nytte af helhedsorienteret viden, der sammenstiller smal viden om effekter med fx viden om, hvorfor noget virker i givne sammenhænge, viden om brugernes oplevelse af indsatsen mv.

En anden opstilling af en evidenstypologi er præsenteret nedenfor. Tankegangen er den, at man bør vælge design efter kendetegn ved henholdsvis studiets (fx en undersøgelse, en evalueringen eller en forskningsoversigt) formål, den pågældende indsats, konteksten for indsatsen og forskerens forhåndsviden om, hvilke virkninger (outcomes) der skal måles, og hvilke årsag-virkningsmekanismer der er på tale. Eksempelvis er RCT velegnet, hvis formålet er at bestemme, om en given indsats bør fortsættes eller nedlægges, hvis indsatsen er af en relativ enkel, standardiseret karakter, hvis konteksten er relativ ensartet (lavt differentieret), og hvis forskerens forhåndsviden om virknings- eller outcome-mål og mekanismer gør det muligt at opstille forholdsvis specifikke hypoteser. Derimod bør fx casestudier vælges, hvis formålet er at tilpasse og forbedre en given indsats og opnå en forståelse af, hvordan indsatsen virker, hvis indsatsen er forholdsvis sammensat og kompleks, hvis konteksten varierer betydeligt, og hvis der er begrænset forhåndsviden om årsag-virknings-processer.

Figur 7.2 En evidenstypologi

	RCT	Forløbs-undersøgelser	Casestudier
Studiets formål	Stop/go af indsats	Stop/go Tilpasning	Forståelse Tilpasning
Indsatsens karakter	Enkel »teknisk«	Veldefineret	Kompleks
Kontekstens karakter	Lavt differentieret	Moderat differentieret	Højt differentieret
Forhåndsviden om årsag-virkning	Hypoteser om specifikke virkninger	Modelleres statistisk	Lille viden

Evidenstypologitænkning kan karakteriseres som en situationsbestemt tilgang til anbefalinger om undersøgelsesdesign. Da vi i denne rapport har lagt stor vægt på at diskutere RCT, vil vi afslutningsvis opsummere de betingelser, hvorunder anvendelse af RCT er særlig relevant.

7.1 Betingelser, hvor RCT er særligt relevant

Frem for urefleksivt at tage afsæt i evidenshierarkiet er det efter vores opfattelse mere konstruktivt at tænke undersøgelsesdesign situationsbestemt. I denne sammenhæng kan vi tale om, at en række betingelser bør være til stede for, at RCT-designet er relevant. Konkret drejer det sig om følgende betingelser (se fx EvalTalk 2004):

- Der er tale om en afgrænset, specificeret intervention.
- Interventionens gennemførelse kan standardiseres.
- Gennemførelsen kan kontrolleres.
- Der foreligger gyldige og pålidelige mål for outcome.
- Randomisering (tilfældig udvælgelse til hhv. kontrol- og interventionsgruppe er mulig).
- Randomisering er etisk forsvarlig.

Eksempler på interventioner, hvor disse betingelser ofte vil være til stede, er forsøg med lægemidler, forsøg med gødningsdosering i landbrug, enkelte sundhedsinterventioner, som fx tandbørstning.

I en række tilfælde er disse betingelser imidlertid ikke til stede. Her kan nævnes sammensatte og kontekstspecifikke forsøg med udvikling af bysamfund og lokalsamfund, og mange typer af sociale interventioner. Her kan man i stedet gennemføre fx casestudier, sammenlignende analyser af

naturlige variationer, anvende triangulering af flere datakilder og metoder, etnografiske feltstudier, registeranalyser af udvikling over tid etc.

I evalueringsteori er det almindeligt anerkendt, at interventionens grad af udvikling også er afgørende for, hvilke evalueringsdesign man bør vælge. En intervention, der er ved at blive udformet, kræver andre evalueringsdesign end en intervention, der har mange år på bagen, og er på et fuldt udviklet og modent stadie (Chen 2005). Det indebærer, at RCT ikke kan vælges generelt, men at interventionens udviklingsstade er én betingelse, man skal være opmærksom på.

Inden for lægemiddelafprøvning, hvor RCT er anvendt, hører RCT da også til en bestemt fase i afprøvningsprocessens fire faser (Trochim, Cornell University, EvalTalk, Dec 4, 2003):

- Fase 1: små eksplorative studier af få cases
- Fase 2: statistiske studier baseret på statistiske korrelationsanalyser
- Fase 3: Først i denne fase gennemføres RCT
- Fase 4: udforskning af dose-respons, mulige utilsigtede effekter, generaliserbarhed

Alle fire faser betragtes som nødvendige for at foretage videnskabelige vurderinger i lægevidenskaben. Fase 3 studier med RCT spiller en væsentlig rolle i sikringen af intern validitet, men tillades normalt kun, når fase 1 og 2 studier er gennemført (som kan tage flere år). De øvrige faser belyser vigtige emner som konstruktions- og ekstern validitet. Typisk vurderes lidt under halvdelen af alle interventioner (i medicin) som sikre nok til, at man går videre til fase 3. Se eksempel om kræftbehandling: www.cancer.gov/clinicaltrials/resources/clinical-trials-education-series (opslag 2/2/2007).

Som det er fremgået, kan der modsætningsvis argumenters for, at når disse betingelser for RCT ikke er til stede, er andre former for design og fremgangsmåder mere relevante. Inden for evalueringstraditionen har man lanceret en såkaldt »contribution analysis«, som et alternativ til RCT til at håndtere det såkaldte »attribution«-problem, dvs. i hvilke udstrækning kan effekter tilskrives indsatsen.

7.2 Et alternativ: Tilvækstanalyse

Tilvækstanalyse (på engelsk: contribution analysis) er helt generelt en fremgangsmåde inden for evaluering, hvor man gennem andre metoder og design end RCT kan demonstrere, at en indsats har gjort en forskel i det felt, man har villet påvirke. Fremgangsmåden har ikke til formål at bevise i RCT-forstand, at indsatsen har bidraget med en tilvækst (et positivt outcome) – eller ej, men at fremlægger data, der formindsker usikkerheden om, hvilken tilvækst indsatsen eventuelt har ydet (Mayne 1999 og Mayne 2001). Tankegangen i tilvækstanalyse bygger på, at der er andre kausalitetsbegreber og -logikker end dem, der knytter sig til RCT ræsonnement om det kontrafaktiske problem. Inden for den empiriske-analytiske videnskabstradition betyder kausalitet, at »hvis X hænder (som fx en indsats), så vil der blive observeret et resultat Y«. Beviset på, at en sådan årsag-virkningsrelation eksisterer, skulle ifølge denne tankegang etableres gennem det kontrafaktiske, nemlig ved at udforme fremgangsmåder, der demonstrerer at uden X, intet Y. Dette gøres ifølge denne tankegang mest ideelt gennem RCT eller ved logikkens afspejling i statistiske multivariate analyser, hvor andre mulige årsagsvariabler »holdes konstant«. Kausalitetsbegrebet har imidlertid andre udlægninger, og der er peget på andre måder at udlede kausalitet på. Her kan kort nævnes Pawson (2006), som skelner mellem forskellige modeller (eller begreber) for kausalitet: rækkefølge (sequential) kausalitet, »konfigurativ« kausalitet og »generativ« kausalitet. Det sidstnævnte begreb bygger på Pawson og Tilley's (1997) context, mechanisme og outcome (CMO) model, som er forankret i en realistisk videnskabstradition. Også Patton har diskuteret alternative måde for at vurdere kausalitet end RCT (Patton 2006).

Tilvækstanalyse, som oprindelig stammer fra finansiel vurdering af foretningsområder, er blevet taget op i en tilpasset form inden for evalueringsområdet (Hill og Cardno 2006), hvor Mayne (2007) giver denne definition:

Contribution analysis is based on the existence of or, more usually, the development of a postulated theory of change for the program being examined. A theory of change sets out why it is believed that the program's activities will lead to a contribution to the intended results; that is, why

and in what manner the observed results can be attributed to the program. The analysis tests this theory against logic and the evidence available on the various assumptions behind the theory of change, and examines other influencing factors. The analysis either confirms (verifies) the postulated theory of change or suggests revisions in the theory where the reality appears otherwise. The analysis is best done iteratively, building up over time a more robust contribution story. The overall aim is to reduce the uncertainty about the contribution the program is making to the observed results through an increased understanding of why the observed results have occurred (or not) and the roles played by the program and other factors.

Vægten i denne analysemåde ligger på at udvikle programteori, som grundlag for empiriske analyser, og herigennem at reducere usikkerhed om indsatsens effekt ved etablering af sandsynlige sammenhænge (plausible associations). Der konstrueres på dette grundlag en troværdig tilvækst-fortælling (contribution story). En sådan fortælling indeholder den indsamlede viden og beskriver indsatsens kontekst, kæder hvorigennem indsatsen fører til effekter, sikring af kvalitet af data og analyse og en diskussion af alternative forklaringer på evalueringens resultat.

I Hill og Cardno (2007) gives et eksempel på en sådan fortælling på bistandsområdet (uddannelse).

8 Konklusion, diskussion og perspektivering

Vi har i det foregående på så vidt mulig neutral vis beskrevet først retningslinjer for udvælgelse af primærstudier til systematiske forskningsoversigter, som de fremgår af de evidensproducerende organisationers hjemmesider (ultimo 2006). Dernæst har vi redegjort for den nyere diskussion om evidenshierarkiet med vægt på argumenter for og imod RCT. Vi har opsummeret begrundelserne for RCT som toppen af hierarkiet og søgt at gruppere de forskellige kritikker heraf. I forlængelse af kritikken og for at sætte kritikken i perspektiv har vi omtalt forskellige alternative undersøgelsesdesign til at forstå og afdække årsags-virknings-relationer. Vi har endvidere berørt de retningslinjer, der ligger for at vurdere kvaliteten af primærstudier, der er gennemført med et givet design, fx kriterier for at vurdere, om RCT er godt eller dårligt gennemført.

Men hvordan ser vi status nu, og hvad mener vi selv?

1. Debatten om evidenshierarkiet, for og imod RCT, og hvor snittet bør lægges, er især på nogle områder og i visse lande præget af næsten videnskabsideologiske forskansninger – samtidig med at der er tegn på mere nuancerede holdninger til RCT og andre design. En situationsbetinget teori om anvendelse af forskningsdesign er under udvikling. Debatten kommer imidlertid sjældent ret dybt ned i de videnskabsparadigmatiske forudsætninger for dette eller hint design, men holder sig oftest på et praktisk-metodologisk niveau.

2. Fortalere for evidenshierarkiet og RCT har forskellig dominans på forskellige politikområder. På sundhedsområdet er evidenshierarkiet og

RCT dominerende, selv om der på forebyggelsesområdet er en opblødning. På det kriminologiske område synes evidenshierarkiet og RCT at have betydelig vægt. På det sociale område og på uddannelsesområdet pågår en levende pro/contra diskussion. I USA synes RCT-dominansen i vækst, mens dette ikke i samme omfang er tilfældet i Europa – er vores skøn.

3. Vi mener, at disse forskelle mellem policy-områder kun i beskeden grad lader sig forklare ud fra indholdet i den pågældende sektors ydelser/opgaver. Forklaring skal snarere søges i forskelle i fagprofessionelle traditioner og -interesser samt i forskellige videnskabsteoretiske paradigmer i de discipliner, der danner baggrund for den fagprofessionelle praksis. Der er dog også forskellige opfattelser inden for samme sektor, som antagelig skal søges i de multiparadigmatiske videnskaber, der ligger til grund for den fagprofessionelle praksis. Strukturelle positioner og interessevaretagelse (autonomi, vidensautoritet, kontrol) spille sandsynligvis også en rolle.

4. Vi finder mange ligheder mellem diskussionen i og om evidensbevægelsen og tidligere metodedebatter inden for samfundsvidenskaberne, fx om videnskabsteoretiske paradigmer og kvalitative versus kvantitative metoder. En debat, som også har afspejlet sig i evalueringslitteraturen over flere årtier.

5. Det nye i metodedebatten i evidensbevægelsen er, efter vores opfattelse, ikke knyttet til indholdet i diskussionen, men til det forhold, at metodediskussionerne er flyttet uden for forskningskredse til det politiske og fagprofessionelle niveau, og dermed inddrager langt bredere og mere heterogene grupperinger end tidligere.

6. Vi forfægter en opfattelse a la evidenstypologien, og at forskningskvalitet bør vurderes ud fra andre kriterier end dem, der fremgår af en rangorden af design (evidenshierarkiet). Hvert design og hver metode bør vurderes på egne præmisser. Flere undersøgelsesdesign kan med

fordel kombineres i ambitionen om at generere mere helhedsorienteret viden om indsatser.

Vi er skeptiske over for værdien af at definere evidensbasen smalt. Brugerne har efter vores erfaring mere nytte af et mere helhedsorienteret perspektiv på både primærstudier og systematiske reviews. Evidensbevægelsen bør organiseres ikke kun med afsæt i RCT-forankrede præmisser, ej heller kun med afsæt i kvantitiativ metodologi.

Bilag 1

Oversigt over vejledninger fra evidensproducerende organisationer

Kommentar til matrixen

Matrixen er struktureret sådan, at første række (helt hvid) indeholder overordnende oplysninger om organisationen. Dernæst kommer rækker med de vejledninger o.a., som organisationerne har udgivet.

I flere vejledninger optræder de samme begreber. Da der kan være forskelle i, hvordan de enkelte organisationer definerer begreberne, er disse uddybet, hvor det er fundet relevant. Ellers henvises til forklaringer i andre celler. De nævnte krydshenvisninger begrænser sig til de organisationer, der er inkluderet i matrixen.

Tegnbrug:
Et tal i en sort cirkel [❶] i kolonnen »terminologi for kvalitetsvurdering, samt vurderingsværktøjer« henviser til det samme tal i kolonnen »kommentar«, hvor der vil være en uddybning.
[-] betyder: irrelevant at udfylde celle.
[□] betyder: informationen mangler.
[▸] betyder: findes på følgende sted i vejledning.

Oversigt over vejledninger o.l. fra udvalgte evidensproducerende organisationer

Organisation	Vejledningens fulde titel	Udgivelses-tidspunkt	Omfang	Design-rangordning	Kvalitetsvurdering af enkelte design	Terminologi for kvalitetsvurdering, samt vurderingsværktøjer	Kommentar	Krydshenvisninger
The Cochrane Collaboration	-	-	-	Anvendelsen af RCT prioriteres højest. Kvalitative design ses som støttende eller underbyggende.	-	-	Bredt perspektiv; dog hovedfokus på effektstudier.	Henviser til CRD´s guideline.
	Cochrane Handbook for Systematic Reviews of Interventions ver. 4.2.6[1]	September 2006	257	Anvendelsen af RCT prioriteres højest. Kvalitative design ses som støttende eller underbyggende.	Ja, iht. bedste design for effektstudier; RCT optræder således som guldstandard	❶ Validitet. ❷ Bias. (► side 80) ❸ Vurderingsværktøjer nævnes. (► side 83)	Bredt perspektiv; dog hovedfokus på effektstudier. ❶ I hvilken grad et studies design og gennemførsel forebygger systematiske fejl (bias) ❷ Systematiske fejl. Fire mulige kilder til bias nævnes: ○ Selection bias. Systematisk forskel i sammenligningsgrupperne. ○ Performance bias. Systematisk forskel i plejen, ud over den intervention, som ønskes evalueret. ○ Attrition bias. Systematisk forskel i de der falder fra i undersøgelsesforløbet. ○ Detection bias. Systematisk forskel i vurderingen af outcome ❸ Der henvises til Moher et al. som har lavet en oversigt over skalaer og tjeklister som kan anvendes til vurdering af RCT validitet (Moher et al. Assessing	Der henvises til litteratur fra CRD og The Campbell Collaboration

Organisation	Vejledningens fulde titel	Udgivelses-tidspunkt	Omfang	Design-rangordning	Kvalitetsvurdering af enkelte design	Terminologi for kvalitetsvurdering, samt vurderingsværktøjer	Kommentar	Krydshenvisninger
							the quality of randomized controlled trials: an annotated bibliography of scales and checklists. Controlled Clin Trials 1995; 16:62-73).	
The Nordic Cochrane Centre	-	-	-	-	-	-	Bredt perspektiv; dog hovedfokus på effektstudier.	Links til The Cochrane Collaboration.
Centre for Reviews and Dissemination (CRD)	-	-	-	Anvendelsen af RCT prioriteres højest. Kvalitative design ses som støttende eller underbyggende.	-	-	Bredt perspektiv; dog hovedfokus på effektstudier.	-
	Undertaking Systematic Reviews of Research on Effectiveness[2]	Marts 2001	ca. 164	Anvendelsen af RCT prioriteres højest. Kvalitative design ses som støttende eller underbyggende.	Kriterier opstilles for effektivitetsstudier, test-accuracy-studier, kvalitative studier samt økonomiske evalueringer (►Stage II, Phase 5, side 8)	❶ Studiekvalitet (svarer til Cochranes "validitet") ❷ Bias. ❸ Intern validitet. ❹ Ekstern validitet. ❺ Vurderingssværktøjer nævnes. (► Stage II, Phase 5 side 6)	Bredt perspektiv; dog hovedfokus på effektstudier. ❶ Graden af, hvor godt et studie inkorporerer tiltag der minimerer bias og dermed styrker den interne validitet. ❷ Tendensen til at producere resultater, der systematisk afviger fra det "sande" resultat. Ikke-biased resultater er internt valide. Fire biastyper nævnes: ○ Selection bias. ○ Performance bias. ○ Attrition bias. ○ Measurement bias (kaldes "Detection bias hos Cochrane). ❸ Hvor sandsynligt det er, at studiets resultater er tæt på "sandheden". Intern validitet er en forudsætning for ekstern validitet	Henviser til The Cochranes Collaborations Håndbog.

Organisation	Vejledningens fulde titel	Udgivelses-tidspunkt	Omfang	Design-rangordning	Kvalitetsvurdering af enkelte design	Terminologi for kvalitetsvurdering, samt vurderingsværktøjer	Kommentar	Krydshenvisninger
							❹ Graden af hvor anvendelige de observerede effekter er uden for studiet. ❺ ○ Individuelle kvalitetskomponenter som metodologi, skjult allokation, blinding, follow-up etc. ○ Kvalitetstjeklister. Lister som ikke scores nummerisk. ○ Kvalitetsskalaer. Numerisk scoring som giver kvantitativt estimat af studiekvalitet. ○ Der advares mod brug af vurderingsværktøjer, uden at være strengt bevidst om disses begrænsninger	
Nordisk Campbell Center (NC2)	-	-	-	Anvendelsen af RCT prioriteres højest. Kvalitative design ses som støttende eller underbyggende.	-	-	Ved forespørgsel (pr. mail) om vejledninger, svarer NC2, at de ikke har en officiel håndbog, men at der efter sigende er en i proces. På hjemmesiden findes dog retningslinjer for, hvordan udarbejdelse af systematiske reviews bør foregå (se http://www.sfi.dk/sw43740.asp), herunder følgende vejledning:	The Campell Collaboration samt The Cochrane Collaborations Handbook.
	How to make a Campbell Collaboration Review	□	217	-	-	-	Vejledningen er opdelt i tre dokumenter, som er beskrevet i det følgende:	-

Organisation	Vejledningens fulde titel	Udgivelses-tidspunkt	Omfang	Design-rangordning	Kvalitetsvurdering af enkelte design	Terminologi for kvalitetsvurdering, samt vurderingsværktøjer	Kommentar	Krydshenvisninger
	Title Registration[3] **(vejledning 1)**	□	40	Nej.	Nej.	Nej.	Indeholder: O Beskrivelse af Campbell Collaboration og NC2. O Beskrivelse af, hvad et Campbell review er. O Beskrivelse af hvordan Reveiw Teams fungerer. O Titelregistreringsformular. O Vejledning i udfyldning af titelregistreringsformularen og en tjekliste. O Eksempel på titelregistrering. O Henvisninger til videre læsning.	The Cochrane Collaboration
	The Protocol[4] **(vejledning 2)**	□	42	Nej.	Nej.	Nej.	Indeholder: O Hvordan man laver en NC2-protokol. O Eksempel på en NC2-protokol. O Protokoltjekliste for reviewere. O Henvisning til mere information. O Udfyldningsskema for tidsforbrug ved udarbejdelse af systemetiske reviews.	The Cochrane Collaboration, CRD
	The Review[5] **(Vejledning 3)**	□	135	Designrangorden diskuteres. Anvendelsen af RCT prioriteres dog højest. Kvalitative design ses som støttende eller underbyggende.	-	-	Indeholder: O Hvordan man laver et C2-review, herunder: ○ Reseach Design Policy Brief (uddybes i det følgende) ○ Statistical Analysis Policy Brief.	The Cochrane Collaboration, CRD

Organisation	Vejledningens fulde titel	Udgivelses-tidspunkt	Omfang	Design-rangordning	Kvalitetsvurdering af enkelte design	Terminologi for kvalitetsvurdering, samt vurderingsværktøjer	Kommentar	Krydshenvisninger
							○ Information Retrieval Policy Brief. ○ Tjekliste for reviewere. ○ Henvisning til videre læsning.	
The Campbell Collaboration (C2)	-	-	-	Anvendelsen af RCT prioriteres højest. Kvalitative design ses som støttende eller underbyggende.	-	-	På hjemmesiden ligger diverse dokumenter som enten er opsummeringer af, eller delelementer i "How to make a Campbell Collaboration Review", som kan fundes på NC2´s hjemmeside. I det følgende gennemgås først Research Design Policy Brief fra ovenstående "Vejledning 3 – The Review", idet dette er et centralt dokument. Dernæst gennemgås resterende relevante dokumenter på C2´s hjemmeside.	The Cochrane Collaboration, What Works Clearinghouse.
	Research Design Policy Brief[6]	November 2004	25	Designrangorden diskuteres. Anvendelsen af RCT prioriteres dog højest. Kvalitative design ses som støttende eller underbyggende.	Kvalitetsvurdering diskuteres	❶ Validitet. ❷ Vurderingsværktøjer diskuteres.	Diskussionsdokument om, hvilke studietyper Campbells database bør indeholde. Hovedspørgsmålet er, hvad C2´s politik skal være iht. hvilke metodologier der er acceptable, når primærstudier skal inkorporeres i systematiske oversigter, der har med interventionseffekt at gøre. Herudover indeholder dokumentet forslag til kodning af forskellige forskningsdesign. ❶ I hvilken grad et studies	Ingen

Organisation	Vejledningens fulde titel	Udgivelses-tidspunkt	Omfang	Design-rangordning	Kvalitetsvurdering af enkelte design	Terminologi for kvalitetsvurdering, samt vurderingsværktøjer	Kommentar	Krydshenvisninger
							design og gennemførsel forebygger systematiske fejl (bias) ❷ Der advares mod brug af vurderingsværktøjer uden at være klar over deres begrænsninger og det foreslås at studiers konklusioner eller delkonklusioner vurderes enkeltvis i stedet for at blive summeret til en samlet score.	
	Guidelines for the Preparation of Review Protocols ver. 1.0[7]	Januar 2001	9	Anvendelsen af RCT prioriteres højest. Kvalitative studier ses som bidrag til: forståelse af intervention, outcomemål, udarbejdelse af forskningsspørgsmål, samt til forståelse effektstudiers heterogenitet.	Nej.	Nej.	Overordnet beskrivelse af gennemførsel af review-protokol, overordende punkter er: O Rationale for protokol O Indhold af en protokol O Registrering af protokol O Ændringer i protokollen O Citation O Kilder der kan assistere i protokolskrivningen	The Cochrane Collaboration, CRD
	Steps in Proposing, Preparing, Submitting and Editing of Campbell Collaboration Systematic Reviews[8]	□	4	Nej.	Nej.	Nej.	Kort overordnet beskrivelse af hvordan systematiske reviews foreslås, forberedes, forelægges og redigeres. Til dokumentet er tilføjet et flowchart for udviklingen af nye systematiske reviews[9].	Ingen.
The Evidence for Policy and Practice In-	-	-	-	Nej	-	-	Metodepluralistisk tilgang.	The Cochrane Cochrane, The Campbell Collabora-

Organisation	Vejledningens fulde titel	Udgivelses-tidspunkt	Omfang	Design-rangordning	Kvalitetsvurdering af enkelte design	Terminologi for kvalitetsvurdering, samt vurderingsværktøjer	Kommentar	Krydshenvisninger
formation and Coordinating Centre (EPPI)								tion
	EPPI-Centre methods for conducting systematic reviews [10]	March 2007	18	Nej.	Ja.	Det fremføres, at vurderingen af studier bør baseres på fire kriterier, navnlig: ❶ Metodologisk kvalitet. ❷ Metodologisk relevans. ❸ Emnerelevans. ❹ Overordnet vurdering.	Vejledningen er en overordnet metodeguide. Indeholder kapitler bl.a. om: O Hvordan man kommer i gang. O Hvordan man indsamler og beskriver forskning på et område. O Hvordan man vurderer og syntetiserer data. O Hvordan man skriver et systematisk review. O Hvordan man anvender og vedligeholder et systematisk review. ❶ Resultaternes pålidelighed, vurderet efter normer for studiekvalitet mht. gennemførsel af det specifikke design. ❷ Graden af hvor passende det specifikke forskningsdesign er for at besvare forskningsspørgsmålet i den systematiske oversigt. ❸ I hvilken grad fokus er passende for at besvare reviewspørgsmålet. ❹ Kaldes WoE – Weigth of Evidence. Bedømmelse af den samlede evidenstyngde på baggrund af vurderingen fra de tre ovenstående kriterier.	Ingen.

Organisation	Vejledningens fulde titel	Udgivelses-tidspunkt	Omfang	Design-rangordning	Kvalitetsvurdering af enkelte design	Terminologi for kvalitetsvurdering, samt vurderingsværktøjer	Kommentar	Krydshenvisninger
	Bog: Using Research For Effective Health Promotion	August 2001	176	□	□	□	Fremføres på hjemmesiden som en beskrivelse af EPPI´s forskningsmetoder. Diskuterer skellet mellem forskning og praksis og argumenterer for, at [health promotion]-service bør baseres på empirisk forskning.	□
	Guidelines for the REPOrting of primary empirical research Studies in Education (The REPOSE Guidelines)11	2004		Nej.	Ja.	Værktøj, der anviser, hvordan primærstudier bør præsenteres.	Rapporten gennemgår ,hvordan empirisk forskning inden for uddannelse afrapporteres. Videre fremføres råd til producenter og læsere om, hvilke aspekter der bør fokuseres på iht. forberedelse, gennemførelse og læsning af primærstudier. Gennemgår hovedovervejelser som: O Studiets introduktion (rationale). O Studiets metoder. O Studiets resultater. O Studiets konklusioner.	Ingen.
	Guidelines for the REPOrting of primary empirical research Studies in Education (The REPOSE Guidelines) Draft for consultation[12]	November 2005	1	Nej.	Ja.	Værktøj, der anviser, hvordan primærstudier bør præsenteres.	Resumé af ovenstående vejledning.	Ingen.

Organisation	Vejledningens fulde titel	Udgivelses-tidspunkt	Omfang	Design-rangordning	Kvalitetsvurdering af enkelte design	Terminologi for kvalitetsvurdering, samt vurderingsværktøjer	Kommentar	Krydshenvisninger
	EPPI-Centre Keywording Strategy for classifying educational research ver. 0.9.713	□	18	Nej.	Nej.	Nej.	Guideline til anvendelse af nøgleord ifm. kategorisering af studier.	Ingen.
	EPPI-Centre Educational Keywording Sheet.[14]	□	1	Nej.	Nej.	Nej.	Oversigtsark ifm. ovenstående keywordingstrategy.	Ingen.
Social Care Institute for Excellence (SCIE)	-	-	-	Nej.	-	-	Metodepluralistisk tilgang. Ingen decideret vejledning pt. Der arbejdes dog på at opdatere en vejledning fra 2001, men den er ikke udkommet endnu. Der findes dog to rapporter, som er relevante for forståelsen af organisationens arbejdsmetoder mht. integration og systematisering af forskning, navnlig:	-
	Using evidence from diverse research designs. SCIE reports no. 3.[15]	November 2003	42	Nej.	Nej.	Nej.	Rapporten giver et overblik over, hvilket arbejde der udføres i diverse institutioner mht. til at udvilkling af vejledninger, manualer ao., der har til formål at udvikle metoder til syntetisering af evidens fra forskellige kvalitative eller miksede forskningsdesigns.	The Cochrane Collaboration, The Campbell Collaboration.
	Knowledge review 3. Types	November 2003	97	Nej.	Ja.	❶ TAPUPAS ❷ **T**ransparency	Rapporten udspringer fra et otte-måneders projekt	The Cochrane Collaboration,

Organisation	Vejledningens fulde titel	Udgivelses-tidspunkt	Omfang	Design-rangordning	Kvalitetsvurdering af enkelte design	Terminologi for kvalitetsvurdering, samt vurderingsværktøjer	Kommentar	Krydshenvisninger
	and quality of knowledge in social care[16]					❸ **A**ccuracy ❹ **P**urposivity ❺ **U**lility ❻ **P**ropriety ❼ **A**ccessibility ❽ **S**pecificity	der skulle udforske typen og kvaliteten af viden inden for socialområdet og hvordan praktikere bør forholde sig til disse former for viden. Generelle konklusioner og åbne spørgsmål er: ○ Spørg om den specifikke viden er TAPUPAS ○ Hver type videnskilde bør lære af de standarder der anvendes i andre videnskilder. ○ Ingen standardskabelon kan erstatte vurdering af kvalitet. ○ Hvem bestemmer ,hvad for forskning der igangsættes? ○ Hvilken værdi kan sættes på personlig erfaring? ○ Hvis mening tæller mht. hvad der er vigtige outcome? ❶ Forkortelse over række spørgsmål, der bør stilles ifm. vurderingen af enhver form for viden. ❷ Er grunden for den specifikke viden og baggrunden for den klar? ❸ Er det ærligt velfundderet på relevant viden? ❹ Er de anvendte metoder egnet til formålet? ❺ Er det anvendeligt? Besvarer det de spørgsmål	The Campbell Collaboration.

Organisation	Vejledningens fulde titel	Udgivelses-tidspunkt	Omfang	Design-rangordning	Kvalitetsvurdering af enkelte design	Terminologi for kvalitetsvurdering, samt vurderingsværktøjer	Kommentar	Krydshenvisninger
							det stiller? ❻ Er det lovligt og etisk forsvarligt? ❼ Er det almenforståeligt? ❽ Lever det op til de standarder, der gælder for denne type viden?	
National Institure for Health and Clinical Excellence (NICE)	-	-	-	-	-	-	NICE har meget støttelitteratur omhandlende evidensbaseret klinisk praksis. Hovedudgivelsen er (se næste række):	The Cochrane Collaboration.
	The guidelines manual 2006[17]	2006	190	EBM-hierarki	Ja, udfyldningsskema for hver studietype.	○ Intern validitet ○ Ekstern validitet ○ Bias ❶ Vurderingssværktøjer nævnes. (► Chapter 7, side 4)	Omfattende vejledning, der forklarer, hvordan NICE udvikler kliniske guidelines og giver råd til tekniske aspekter af guidelineudvikling. ❶ AGREE-Instrument. Til vurdering af metodologisk kvalitet af kliniske vejledninger. (http://www.agreecollaboration.org/pdf/dk.pdf)	The Cochranes Collaborations Håndbog, CRD´s håndbog.
Institutet för utveckling av metoder i socialt arbete (IMS)	-	-	-	-	-	-	-	The Campbell Collaboration.
What Works Clearinghouse (WWC)				Ja.	Ja, iht. RCT.		Udelukkende fokus på effektstudier.	Samarbejder med og henviser til The Campbell Collaboration.
							Ved forespørgsel (pr. mail) om guidelines svarer IMS, at de ikke har en guideline, men at foreligger et førsteudkast til en	The Cochranes Collaborations Håndbog samt deres metaanalyseprogram RevMan.

Organisation	Vejledningens fulde titel	Udgivelses-tidspunkt	Omfang	Design-rangordning	Kvalitetsvurdering af enkelte design	Terminologi for kvalitetsvurdering, samt vurderingsværktøjer	Kommentar	Krydshenvisninger
							sådan.	
	Evidence Standards for Reviewing Studies[18]	September 2006	12	Ja.	Ja, iht. RCT.	❶ Meets Evidence Standards ❷ Meets Evidence Standards with Reservations ❸ Does not meet evidence Standards	I første omgang screenes studier iht. relevans ift. emneområde, kvaliteten af outcomemål, samt om data er blevet afrapporteret tilstrækkeligt. Accepterede studier vurderes iht. i hvilken grad de bidrager med evidens for den testede interventions effektivitet. ❶ Strong evidence: Veldesignede og implementerede RCT eller gode quasi-eksperimenter ❷ Weaker evidence: Quasieksperimenter uden større design- eller implementeringsvanskeligheder eller RCT med større design- eller implementeringsvanskeligheder. ❸ Insufficient evidence. Studier, der ikke lever op til ovenstående krav	

1 http://www.cochrane.org/resources/handbook/Handbook4.2.6Sep2006.pdf
2 http://www.york.ac.uk/inst/crd/report4.htm
3 http://www.sfi.dk/graphics/Campbell/Dokumenter/For_Forskere/guide_1_title_samlet20DEC04.pdf
4 http://www.sfi.dk/graphics/Campbell/Dokumenter/For_Forskere/Guide_2_protocol_dec04.pdf
5 http://www.sfi.dk/graphics/Campbell/Dokumenter/For_Forskere/guide_3_review_samlet20DEC04.pdf
6 http://www.campbellcollaboration.org/MG/ResDesPolicyBrief.pdf
7 http://www.campbellcollaboration.org/c2_protocol_guidelines%20doc.pdf
8 http://www.campbellcollaboration.org/C2EditingProcess%20doc.pdf
9 http://www.campbellcollaboration.org/guide.flow.pdf
10 http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=89

11 http://www.multilingual-matters.net/erie/018/0201/erie0180201.pdf
12 http://eppi.ioe.ac.uk/cms/Portals/0/PDF%20reviews%20and%20summaries/EPPI%20REPOSE%20Guidelines%20A4%202.1.pdf
13 http://eppi.ioe.ac.uk/EPPIWebContent/downloads/EPPI_Keyword_strategy_0.9.7.pdf
14 http://eppi.ioe.ac.uk/EPPIWebContent/downloads/EPPI_keyword_sheet_0.9.7.pdf
15 http://www.scie.org.uk/publications/reports/report03.pdf
16 http://www.scie.org.uk/publications/details.asp?pubID=31
17 http://www.nice.org.uk/page.aspx?o=308639
18 http://www.nice.org.uk/page.aspx?o=308639

Litteratur

ABAforum (2007): *Evaluering uden evidens. Fagfolk og forskere kommenterer ETIBA-rapporten om autismetilbud*.

Andersen, A.N. og M. Osler (2004): Kohorteundersøgelser for begyndere. *Ugeskrift for Læger*, 15, 1431-1433.

Andersen, Ib (1997): *Den skinbarlige virkelighed*, 1997.

Andersen, L.B. og T.I.A. Sørensen (1997): Design af kohortestudier. *Bibliotek for Læger,* 117-134.

Andersen, Vibeke Normann (2003): Brugerorienteret evaluering. I: Peter Dahler-Larsen m.fl.: *Selvevalueringens hvide sejl*. Odense: Syddansk Universitetsforlag.

Bhaskar's Realist theory of science (1997): *A Realist Theory of Science*, 2nd edn. London: Verso.

Bhatti, Yosef; Hanne Foss Hansen og Olaf Rieper (2006): *Evidensbevægelsens udvikling, organisering og arbejdsform.* En kortlægningsrapport. København: AKF Forlaget.

Bjørn, N.; : Geerdsen og P. Jensen (2004a): *The Threat of Compulsory Participation in Active Labour Market Programmes for Unemployed*. Protocol.

Bjørn, N.; : Geerdsen og P. Jensen (2004b): *The Threat of Compulsory Participation in Active Labour Market Programmes for Unemployed.* Review not yet approved.

Brandt, Preben og Mette Kirk (2003): *Billeder fra hverdagen. En dokumentarisk-analytisk fremstilling af den sociale indsats over for hjemløse, misbrugere eller sindslidende personer, som den ses af de, der arbejder i feltet.* AKF Forlaget.

Bryman, A. (2004): *Social Research Methods*. Oxford: Oxford University Press.

Campbell, D.T. og J.C. Stanley (1966): *Experimental and Quasi-experimental Designs for Research.* Chicago: Rand McNally.

Campbell, Donald T. (1969): »Reforms as experiments«, *American Psychologist* vol. 24, april, pp. 409-429.

Chalmers, I. (2005): If evidence-informed policy works in practice, does it matter if it doesen't work in theory? *Evidence and Policy,* 1(2), 227-242.

Chalmers, Iain (1) (2001): Comparing the like: some historical milestones in the evolution of methods to create unbiased comparison groups in therapeutic experiments, *International Journal of Epidemiology,* vol. 30, pp. 1156-1164.

Chalmers, Iain (2) (2003): Trying to Do more Good than Harm in Policy and Practice: The Role of Rigorous, Transparent, Up-to-Date Evaluations, *Annals of American Political and Social Sciences,* vol. 589.

Chen, Huey-Tsyh (2005): *Practical Program Evaluation. Assessing and Improving Planning. Implementation, and Effectiveness*. Sage Publications. Thousand Oaks. London. New Delhi.

Clarke, A. (2006): Evidence-Based Evaluation in Different Professional Domains: Similarities, Differences and Challenges. I: I.F. Shaw; J.C. Greene and M.M. Mark, *The Sage Handbook of Evaluation*, London: Sage.

Cochrane, Archie L. (1999) [1972]: *Random Reflections on Health Services*, London: Royal Society of Medicine Press Ltd.

Coggon, D.; D. Barker og G. Rose (1997): *Epidemiology for the Uninitiated.* England: British Medical Journal Publishing Group.

Concato, J.; N. Shah, R.I. Horwitz (2000): Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Design. *New England Journal of Medicine,* 342(25), 1887-1892.

Cook, T.D. (2002): Randomized Experiments in Educational Policy Research: A Critical Examination of the Reasons the Educational Evaluation Community has Offered for not Doing Them. *Educational Evaluation and Policy Analysis*, Fall, Vol. 24(3): 175-199.

Cook, T.D. (2003): *A Critical Appraisal of the Case Against Using Experiments to Assess School (or Community) Effects*, 128 kB pdf at <http://www.educationnext.org/unabridged/20013/cook.pdf> Print version: Why have educational evaluators chosen not to do randomized experiments? *Annals of American Academy of Political and Social Science*, 589: 114-149.

Cook, Thomas D. (2003): Why have Educational Evaluators Chosen Not to Do Randomized Experiments. *Annals of the American Academy of Political and Social Science,* vol. 589.

Cumming, P. og N.S. Weiss (1998): Case series and exposure series: the role of studies without controls in providing information about the etiology of injury or disease. *Injury Prevention,* 4, 54-57.

Dahler-Larsen, Peter & Hanne K. Krogstrup (2003): *Nye Veje i Evaluering*. Viborg: Systime.

Dart, J. og J. Mayne (2004): Performance Story. I: S. Mathison (Ed.): *Encyclopedia of Evaluation*. Thousand Oaks, CA: Sage Publications.

Davies, T.O.; S. Nutley og N. Tilley (2004): Debates on the role of experimentation. I: Davies, T.O.; S. Nutley & P.C. Smith: *What Works? Evidence-based policy and practice in public services*. Bristol: The Policy Press (251-275).

Day, S.J. og D.G. Altman (2000): Blinding in clinical trials and other studies. *BMJ,* 321, 504.

Deahl, M. (2006): Comments to the review. I: S. Rose; J. Bisson, R. Churchill and S. Wessely: Psychological debriefing for preventing post traumatic stress disorder (PTSD), *Cochrane Database of Systematic Reviews* Issue 2, Art.

Ekeland, Tor-Johan (2004): *Autonomi og evidensbaseret praksis,* Senter for professionsstudier, Høgskolen i Oslo, Arbeidsnotat nr. 6.

Enheden for Brugerundersøgelser (2007): *Patienters oplevelser på landets sygehuse.* Spørgeskemaundersøgelse blandt 26.045 indlagte patienter 2006. København.

EvalTalk (2004): Archives of EvalTalk, listserv of the American Evaluation Association, January-February, 2004 debate and discussion on AEA position statement on proposed Dept of Education research and evaluation standards.

Farrington D.P. (2003): Methodological Quality Standard for Evaluation Research. *The ANNALS of the American Academy of Political and Social Science*, 587, May (49-68).

Feinstein, Alvan R. og Ralph I. Horwitz (1997): Problems in the »Evidence« og »Evidence based Medicine«, *The American Journal of Medicine,* vol. 103.

Fiona Kotvojs Kurrajong Hill Pty Ltd and Cardno ACIL (2006). Contribution Analysis – A New Approach to Evaluation in International Development. International Conference, the Australian Evaluation Society, Holiday Inn Esplanade, Darwin, Australia, 4-7 September 2006, Final Paper.

Fisker, Jesper; Bo Vinnerljung, Vibeke Jensen & Alice Rasmussen (2007): *Evaluering af Nordisk Campbell Center*. København.

Flyvbjerg, B. (1993): Rationalitet og magt. Bind 1. *Det konkretes videnskab*. København: Akademisk Forlag.

Glendinning, Caroline Sue Clarke; Philippa Hare, Inna Kotchetkova, Jane Maddison & Liz Newbronner (2007): Outcomes-focused services for older people. London: *Social Care Institute for Excellence, Adults' Services Knowledge Review* 13.

Gluud, L.L. (2005): *Bias in Clinical Intervention Research*. Doktordisputats, ni publicerede artikler, samt samlet oversigt. København: Det Sundhedsvidenskabelige Fakultet, København Universitet.

Gustavsen, B. (2001): Theory and Practice: The mediating discourse. I: P. Reason og H. Bradbury (red.). *Handbokk of action research*. London: Sage (17-26).

Gøtzsche, Peter C. (1990): *Bias in double-blind trials* (disp.), Dan Med Bull 1990; 37, pp. 329-36.

Habermas, J. (1968): *Erkenntris und Interesse*. Frankfurt a.M: Suhrkamp.

Hammer, Svein (2004): Hvordan kan vi måle dette? Et essay om tidsriktige evalueringspraksiser. *Sociologi i dag, årgang 34*, nr. 4/2004, s. 9-26.

Hammersley, Martyn (2005): Is the evidence-based practice movement doing more good than harm? Reflections on Iain Chalmers' case for research-based policy making and practice, *Evidence and Practice,* vol. 1, no. 1, pp. 85-100.

Hansen, H.F. (2003): *Evaluering i staten: Kontrol, læring eller forandring*? København: Samfundslitteratur.

Hansen, Hanne Foss & Birte Holst Jørgensen (1995): *Styring af forskning: Kan forskningsindikatorer anvendes?* København: Samfundslitteratur.

Hansen, Hanne Foss & Finn Borum (1999): The Construction and Standardization of Evaluation. The Case of the Danish University Sector. I: *Evaluation*, vol. 5 (3), pp. 303-329.

Hansen, Kasper Møller (2003): Deliberativ demokratisk evaluering. I: Peter Dahler-Larsen m.fl.: *Selvevalueringens hvide sejl*. Odense: Syddansk Universitetsforlag.

Hede, A. (2007): Systematiske reviews og evidens. I: Fuglsang, L.; P. Hagedorn-Rasmussen og P.B. Olsen: *Teknikker i samfundsvidenskaberne*. Frederiksberg: Roskilde Universitetsforlag (28-53).

Henggeler, S.W.; S.K. Schoenwald, C.M. Borduin og C.C. Swenson (2006): The Littel paper: Methodological critique and meta-analysis as Trojan horse, *Children and Youth Services Review*, vol. 28, pp. 447-457.

Higgins, J.T.P. og S. Green (eds.) (2007): Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 (updated September 2006. http://www.cochrane.org/resources/handbook/hbook.htm (besøgt 5. juli 2007).

Høgsbro, Kjeld et al. (2003): *HMS-undersøgelsen*. AKF Forlaget.

Høgsbro, Kjeld (2002): Rehabilitering af mennesker med traumatiske hjerneskader på Kolonien Filadelfia. AKF Forlaget.

Høgsbro, K. (2007): *ETIBA. En forskningsbaseret evaluering af rehabiliterings- og træningsindsatsen for børn med autisme, herunder evaluering af behandlingsmetoden ABA* (Applied Behavior Analysis). Århus: MarselisborgCentret.

Jørgensen, M.H.; M. Riegels, U. Hesse og M. Grønbæk (2006): *Evaluering af forbuddet mod salg af alkohol til personer under 16 år*. København: Statens Institut for Folkesundhed.

Kenardy, J. og V. Carr (1996): Imbalance in the debriefing debate: what we don't know far outweights what we do, *Bulletin of the Australian Psychological Society*, vol. 17, no. 1, pp. 4-6.

Krogstrup, Hanne K. (2006): Evalueringsmodeller. Århus: Academica, 2006 (2. udgave).

Kyriacou, C. og M. Goulding et al. (2004): *A systematic review of the impact of Daily Mathematics Lesson in enhancing pupil confidence and competence in early mathematics*, EPPI-review, December.

Kürstein, P.; J. Kjellberg, L. Herbild, K.R. Olsen, M. Willemann, J. Søgaard og C. Gludd (2005): *Fra forskning til praksis*, Copenhagen: DSI Institut for Sundhedsvæsen.

Launsø, L.; I. Henningsen, J. Rieper, H. Brender, F. Sandø og A. Hvenegaard (2006): Expectatione and effectiveness of medical treatment and classical homeopathic treatment for patients with hypersensitivity illnesses – A one year prospective study. Homeopathy, vol. 96 (1): 233-242.

Launsø, L og N. Haahr (2007): Bridge Building and Integrative Treatment of People with Multiple Sclerosis. Research-basedEvaluation of a Team-building process. Journal of Complementary and Integrative Medicine, vol. 4: 1, Article 7. Available at: http://www.bepress.com/jcim/vol4/iss1/7.

Launsø, L. og O. Rieper (2005): *Forskning om og med mennesker. Forskningstyper og forskningsmetoder i samfundsforskning*. København: Nyt Nordisk Forlag Arnold Busck.

Launsø, Laila og Dorte E. Gannik (2000): The need for revision of medical research designs. In: Gannik, Dorte E. og Laila Launsø (eds.) *Disease, knowledge and society*. Samfundslitteratur, Frederiksberg, pp. 243-261.

Lipsey, Mark W. (1997): What Can You Build With Thousands of Bricks? Musings on the Cumulation of Knowledge in Program Evaluation. *New Directions for Evaluation* 76:7-23.

Littel, J.H. (2005): Lessons from a systematic review of effects of multisystemic therapy, *Children and Youth Services Review*, vol 27, pp. 445-463.

Littel, J.H. (2006): The case for multisystemic therapy: evidence or orthodoxy? *Children and Youth Services Review,* vol. 28, pp. 458-472.

Littel, J.H.; M. Popa og B. Forsythe (2005): *Multisystemic Therapy for Social, Emotional, and Behavioural Problems in Youth Aged 10-17*, Copenhagen: Nordic Campbell Centre.

Madsen, J.S. og I.B. Andersen (2005): At skelne skidt fra kanel – kritisk udvælgelse og læsning af evidens. I: Andersen, I.B. og J.S. Matzen (red.). *Evidensbaseret medicin.* København: Gads Forlag.

Marsh and Fisher (2005): *Developing the evidence base for social work and social care practice*, Report no. 10, London: Social Care Institute for Excellence.

Mayne, John (1999): *Addressing Attribution through Contribution Analysis: Using Performance Measures Sensibly*: Discussion Paper. Office of the Auditor General of Canada. June.

Mayne, John (2001): Addressing Attribution through Contribution Analysis: Using Performance Measures Sensibly. *The Canadian Journal of Program Evaluation,* 16 (1), 1-24.

Mayne, John (2007): Exploring attribution through contribution analysis, draft chapter for a book in progress in the INTEVAL Group (the international evaluation research group).

Melby, J.; O. Rieper og M.Togeby (1993): *Håndbog i evaluering*. AKF Forlaget.

Moher, D.; J. Tetzlaff, A.C. Tricco, M. Sampson og D.G. Altman (2007): Epidemiology and reporting characteristics of systematic reviews. *PLos Med 4 (3): e78. doi: 10.1371/journal.pmed.0040078*.

Moncrieff, J.; R. Churchill, C. Drummond og H. McGuire (2001): Development of a quality assessment instrument for trials of treatments for depression and neurosis, *International Journal of Methods in Psychiatric Research*, vol. 10, no. 3, pp. 126-133.

Moos, Lejf et al. (2005): *Evidens i uddannelse?* København: Danmarks Pædagogiske Universitets Forlag.

Morris, Jenny & Michele Wates (2006): Supporting disabled parents and parents with additional support needs. London: *Social Care Institue for Excellence, Adults' Services Knowledge Review* 11.

Mullen, Edward J. et al. (2004): From concept to implementation: challenges facing evidence-based social work, *Evidence and Practice,* vol. 1, no. 1, pp. 61-84.

Oakley, Ann (2000): A Historical Perspective on the Use of Randomized Trials in Social Science Settings, *Crime and Delinquency,* vol. 46, no. 3, juli 2000, pp. 315-329.

OECD: OECD-U.S. (2004): *Meeting on Evidence-Based Policy Research in Education. Forum Proceedings*. Paris: OECD.

Ogilvie, D.; M. Egan, V. Hamilton og M. Petticrew (2005): Systematic reviews of health effects of social interventions: 2. Best available evidence: how low should you go? *Journal of Epidemiology and Community Health*, 59, 886-892.

Olsen, L. & O. Rieper (2004): Evalueringsbegrebet, modeller og paradigmer. I: O. Rieper (red.) *Håndbog i evaluering*. København: AKF Forlaget (15-33).

Pawson, R. (2006) *Evidence-based Policy. A Realist Perspective*, London: Sage.

Pawson, Ray (2002): Evidence-based Policy: I: search of a method, *Evaluation,* vol. 8, no. 2, pp. 157-181.

Pawson, Ray; Annette Boaz, Lesley Grayson, Andrew Long & Colin Barnes (2003): *Knowledge Review 3. Types and quality of knowledge on social care.* London: SCIE.

Pedersen, Kjeld Møller (2004): *Sundhedspolitik – beslutningsgrundlag, beslutningstagen og beslutninger i sundhedsvæsenet,* Odense: Syddansk Universitetsforlag.

Pedersen, T. et al. (2001): Hvad er evidensbaseret medicin, *Ugeskrift for Læger*, vol. 163, pp. 3769-3772.

Perit, Birger m.fl. (2002): *Evaluering af erfaringerne med institutionsbegrebets ophævelse på handicap-området, 1998-2002*. København: Socialministeriet.

Peters, B.G. (1998): *Comparative Politics, Theory and Method.* New York: University Press.

Petticrew, M. og H. Roberts (2003): Evidence, hierarchies, and typologies: horses for courses, *Journal of Epidemiology and Community Health*, vol 57, pp 527-9.

Petticrew, M. og H. Roberts (2006) *Systematic Reviews in the Social Sciences. A Practical Guide*, Malden: Blackwell Publishing.

Radnitzky, G. (1970): *Contemporary Schools of Metascience*. Göteborg: Akademiförlaget.

Rose, S.; J. Bisson, R. Churchill og S. Wessely (2002): Psychological debriefing for preventing post traumatic stress disorder (PTSD), *Cochrane Database of Systematic Reviews* Issue 2, Art. No.: CD000560. DOI: 10.1002/14651858.CD000560. Available in a commented version from 2006, Issue 4 (Status: *Commented*) Copyright © 2006 The Cochrane Collaboration. Published by John Wiley and Sons, Ltd. DOI: 10.1002/14651858.CD000560:
http://www.mrw.interscience.wiley.com/cochrane/clsysrev/articles/CD000560/frame.html

Rosholm, Michael (2004): Evaluering af aktivering ved hjælp af kvantitative metoder. I: Rieper, Olaf (red.): *Håndbog i evaluering*, ss 224-239, AKF Forlaget.

Sackett, D.L. (1997): Evidence-based medicine. I: Kristensen, F.B. og H. Sigmund (red.): *Evidensbaseret sundhedsvæsen. Rapport fra et symposium om evidensbaseret medicin, planlægning og ledelse.* DSI Institut for Sundhedsvæsen, DSI rapport 97.02.

Sackett, D.L.; S.E. Straus, W.S. Richardson, W. Rosenberg og R.B. Haynes (2000): *Evidence-Based Medicine: How to Practice and Teach EBM.* Edinburgh: Churchill Livingstone.

Sackett, D.L.; W.M.C. Rosenberg, J.A.M. Gray, R.B. Haynes og W. Scot (1996): Evidence based medicine: What it is and what it isn't. *BMJ*, 312(7023), 71-72.

Sauerland, Stefan; Rolf Lefering og E.A.M. Neugebauer (1999): The pros and cons of evidence-based surgery, *Langenbecks Archive of Surgery*, vol. 384, no. 5, pp. 423-431.

Schlosser, R. (2007): Appraising the Quality of systematic Reviews. *Technical Brief, no. 17. National Center for the dissemination of disability Research* (www.ncddr.org/kt/products/focus/focus17/index.html)

Schultz, K.F.; I. Chalmers, R.J. Hayes og D.G. Altman (1995): Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA, 273*(5), 408-412.

Schwartz, Robert og John Mayne (eds) (2005): *Quality Matters*. Transaction Publishers. New Brunswick (USA) and London (UK).

Scocozza, Lone (2000): The randomised trial. A critique from the philosophy of science. I: Gannik, Dorte E. og Laila Launsø (eds.) *Disease, knowledge and society*. Samfundslitteratur, Frederiksberg, pp 231-242.

Sherman, Lawrence W. (2003): Misleading Evidence and Evidence-Led Policy: Making Social Science more Experimental, *The Annals of American Political and Social Sciences,* vol. 589.

Soares-Weiser, K.; M. Brezis, R. Tur-Kaspa og L. Leibovici (2002): *Antibiotic Prophylaxis for Cirrhotic Patients with Gastrointestinal Bleeding*,

the Cochrane Database of Systematic Reviews, Issue 2. Art. No.: CD002907. DOI: 10.1002/14651858. CD002907.

Social Kritik (2005): *Zweifel. Tema: Måling og evidens*, nr. 102.

Stevens, A.; K. Abrams, J. Brazier, R. Fitzpatrick og R. Lilford (eds.) (2001): *Advanced handbook of methods in evidence based healthcare*, London, Thousand Oaks, New Delhi: SAGE Publications.

Trochim (2003): Archives of EvalTalk, listserv of the American Evaluation Association, December.

Unge Pædagoger (2006): Temanummer »*Dansen om evidensen*«, nr. 3, Juli 2006.

Vedung, E. (1998): *Utvärdering i politik och förvaltning*. Lund: Studentlitteratur.

Wampold, Bruce E. (2001): *The Great Psychotherapy Debate: Models, Methods, and Findings* (LEA's Counseling and Psychotherapy Series).

Summary

The Methodology Debate on Evidence

Issued October 2007

by Olaf Rieper and Hanne Foss Hansen

The aim of this report is to provide an overview of central methodological positions and discussions in what is termed the evidence movement. The term »evidence« has become a positive concept in government: evidence-based policy, evidence-based practice, evidence-based management, evidence-based medicine, educational approaches, etc. The essence of the evidence movement is to summarise knowledge from several individual studies and evaluations. The objective is to produce and disseminate the best possible knowledge on the results of given interventions. The evidence movement has gained ground internationally and nationally over the last 10-15 years. We restrict ourselves to what we maintain is the novel aspect of the evidence concept, i.e. the global and international organisations and networks that specialise in producing, commissioning and communicating publicly available systematic reviews to decision-makers in policy and practice. Systematic reviews are abstracts of available research and studies on a given subject, such as the effects of a certain intervention or treatment, and they are carried out in a systematic, transparent way. Our primary focus is on the major welfare areas of healthcare, social issues and education.

Our principal conclusion is that the sometimes lively debate for and against a narrow or broad concept of evidence varies among the different sectors (the healthcare, education and social sectors). The discussions are, to a considerable degree, influenced by the traditions and interests characterising the professional groups in the various sectors. Previously, methodology discussions were mainly conducted among researchers. The evidence movement has moved the methodology discussions beyond the world of research to the political and professional level. The possible consequences are that education, social work and healthcare services may be changed and justified in policy and practice. Research knowledge plays an increasing role in how we design our society, including the public sector. And it seems that the evidence movement can make a significant contribution to this. The danger lies in a narrow definition of the evidence base on the basis of randomised controlled trials and quantitative analyses only. We maintain that the evidence movement should be organised with a broad approach to research methodology.

A number of the organisations and networks that produce and disseminate systematic reviews target certain players in several countries. For example, the Cochrane cooperation operates in the healthcare area, while the Campbell cooperation operates in the social and labour-market areas as well as criminology. Other organisations have national target groups, e.g. the UK-based »Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI)«. We described these organisations in a previous report with a European focus (Bhatti, Hansen and Rieper 2006).

The products of these organisations, i.e. the systematic reviews, have great potential influence on what is accepted by the target groups – decision-makers at various levels – as reliable knowledge. This makes them important generators of knowledge that is regarded as reliable and legitimate for policy and practice. They simply define the boundaries of what can be considered »valid« knowledge. The debate therefore plays a key role in relation to the production of systematic reviews. In other words, the debate is about which methodology should be applied to the preparation of systematic reviews. One particular element of producing systematic reviews is a core issue in the debate: the qualitative assessment of primary studies to decide whether to include them in a given systematic review.

Other methodology issues are also debated, such as methods to synthesise (summarise) the results of multiple primary studies. We will touch on this issue later, but our focus is on the qualitative assessment of primary studies.

In this report we first describe one of the most widely used approaches to this assessment, the «evidence hierarchy«. The »golden standard« of research design, i.e. randomised controlled trials (RCT) tops the hierarchy. Other designs are further down the scale, e.g. longitudinal studies, and, even further down, case studies. As such, the notion is that research design can be ranked to the effect that, assuming optimum implementation, some designs will provide more reliable results than others. We describe the various designs in a typical evidence hierarchy and provide examples of studies based on the various designs.

Secondly, we have reviewed the guidelines and handbooks of 10 evidence-producing organisations in the USA and Europe. Against this background, we have described these organisations' own approaches to review of primary studies. It turns out that six out of the 10 organisations state, in their own guidelines, that they apply the logic of the evidence hierarchy. (In addition, two organisations cite the guidelines of these organisations) The remaining two organisations state that their approaches are not based on a ranking of research design. They write, among other things, that their systematic reviews are not only about the effects of interventions, but also about implementation, user perception, etc. for which designs other than RCT are ideally suited. They apply a broader knowledge base than research, taking into account knowledge acquired by professionals in practice as well as knowledge acquired by users. Furthermore, they integrate research based on different designs into one systematic review. However, a qualitative assessment of primary studies is also required after their design-based selection. After all, like other designs, an RCT may have been carried out more or less successfully. It emerges that, as indicated by the guidelines of the organisations, the organisations that attach importance to the evidence hierarchy apply assessment criteria based on what might be termed a neo-positivistic paradigm with internal validity as its core. On the other hand, the organisations with another point of departure than the evidence hierarchy emphasise the relevance of primary studies to a given

topic and apply a design-based assessment method, as it were. As regards synthesisation of the findings, organisations prioritising RCT (which is at the top of the evidence hierarchy) recommend meta-analysis as the ideal method. The other organisations adopt a pluralistic approach to synthesisation, also recommending narrative and conceptual synthesisation depending on the issue and the design of the primary studies. Some systematic reviews combine several synthesisation methods.

Thirdly, on the basis of examples of systematic reviews and other sources, we present an analysis of evidence-producing organisations' actual compliance with their own methodology guidelines and recommendations. It turns out that the organisations that attach importance to the evidence hierarchy have a greater share of systematic reviews comprising solely RCT-based primary studies. However, even the most ardent advocates of RCT also produce systematic reviews that include quasi experimental and other designs. One explanation for this is that there is simply not sufficient quality RCT available in the relevant area. This applies especially to Europe where RCT-based research in the social and educational sectors is less common than in the USA.

Fourthly, we outline arguments for and against RCT as a research design since RCT, at the top of the evidence hierarchy, is at the core of the methodology debate in and on the evidence movement. The RCT design is suitable for analysing the effects of limited and specific interventions e.g. clinical trials. Randomisation of the intervention and control groups ensures that neither the subjects of the intervention nor the people in charge of the study know who is included in the intervention and control groups, respectively. Furthermore, it is possible to keep all factors constant by establishing baselines before the intervention commences and by measuring the effects of the intervention. The application of RCT also gives rise to challenges and limitations, however. Firstly, RCT designs produce narrow evidence in the sense that they solely have rhetorical force concerning effects, i.e. about which interventions are effective and which are not. They have no rhetorical force concerning why something works or does not work, nor about how the subjects perceive the intervention. Secondly, there are a number of technical problems in some contexts. For example, when applying RCT in the welfare and educational areas, it is often difficult to

ensure blinding, i.e. that the participants in the trial do not know whether they are in the control group or the trial group. In addition, critics have formulated a number of arguments against applying RCT in areas with complex and dynamic interventions where the context influences whether and how the interventions work. The discussions about the strengths and weaknesses of applying RCT disclose variations in understanding of causality and science-theoretical paradigms.

Finally, we briefly introduce the concept of »evidence typology« as an alternative or supplement to the evidence hierarchy. The thinking behind evidence typologies is that different study designs can potentially answer different study questions. The point of departure is not the notion that some study designs are stronger than others, since the challenge lies in adapting the study design to the questions addressed by the study. The typology approach can inspire development of more holistic study designs and systematic reviews, i.e. knowledge about various aspects of a given intervention is assessed on the basis of a whole array of study designs.

Noter

1. Organisationen blev dannet under benævnelsen MTV-Instituttet i 1997, men blev efter en fusion i 1999 omdøbt til CEMTV.

2. Vi retter en stor tak til Yosef Bhatti, Leo Milgrom og Anne Rehder for hjælp til at oparbejde og tjekke datamateriale.

3. http://www.york.ac.uk/inst/crd/

4. http://www.york.ac.uk/inst/crd/report4.htm

Med denne rapport giver vi en oversigt over centrale metodiske positioner og diskussioner inden for det, vi kalder evidensbevægelsen. Vi afgrænser os til det, der efter vores mening er det nye i evidensbegrebet, nemlig at der er etableret globale og nationale organisationer og netværk, der er specialiserede i at producere, bestille og formidle forskningsoversigter til beslutningstagere i politik og praksis – og som er tilgængelig for alle. Forskningsoversigter er sammenfatninger af foreliggende forskning, fx effekter af en bestemt indsats eller behandling, foretaget på systematiske og gennemskuelige måder. Forskningsmæssig viden spiller en stadig større rolle for, hvordan vi former samfundet, herunder den offentlige sektor. Og evidensbevægelsen synes at kunne udgøre et vigtigt bidrag hertil.

AKF
Forlaget

Nyropsgade 37
DK-1602 København V

tel: +45 4333 3400
fax: +45 4333 3401

akf@akf.dk
www.akf.dk

DET SAMFUNDSVIDENSKABELIGE FAKULTET
KØBENHAVNS UNIVERSITET