

Hvordan isolerer vi sammenhængen mellem indsats og resultat? Propensity score matching som metode til effektevaluering

Niels Matti Søndergaard
Metodekonsulent
Danmarks Evalueringsinstitut (EVA)

Rasmus Højbjerg Jacobsen
Seniorrådgiver
Centre for Economic and Business Research (CEBR)
Copenhagen Business School

21. april 2010

Propensity score matching er en relativt ny statistisk metode der blandt andet kan bruges til effektevaluering. Den bærende ide i metoden er at skabe en kontrolgruppe som ligner en bestemt gruppe af borgere eller virksomheder der er målgruppe/genstand for en offentlig indsats, så meget at det er muligt at sammenligne outcomes for de to grupper på udvalgte variable. Artiklen rummer en introduktion til metoden og belyser den gennem en række eksempler. Afslutningsvis diskuteres nogle metodiske udfordringer i forhold til at bruge metoden.

Effektevaluering handler om at undersøge årsagssammenhænge (Munk 2008). Konkret er vi interesserede i at undersøge sammenhængen mellem en given indsats og ændringer i et eller flere outcomes. Det er svært at foretage effektevaluering fordi mange andre forhold end den pågældende indsats påvirker de outcomes vi er interesserede i. Hvis vi f.eks. er interesserede i at undersøge hvordan indførelsen af en mentorordning på erhvervsuddannelserne påvirker elevernes gennemførelses- og frafaldsprocent, må vi være opmærksomme på at udviklingen i frafaldet over en årrække ikke bare påvirkes af den indførte mentorordning, men også af ændringer i konjunkturerne, ungdomskaraktoren og lovgivningen og desuden af mange andre forhold der er svære at sætte på en formel.

Hvordan man forholder sig til dette problem, er der forskellige bud på. Inden for evalueringsfeltet i Danmark har der især været diskussion af forskellige evalueringsmodeller der baserer sig på indsatsteori som værktøj til at analysere de processer en given indsats igangsætter, og de effekter eller bidrag til effekter den igangsætter. Diskussionerne har især taget udgangspunkt i evalueringsmodeller som virkningsevaluering, realistic evaluation og contribution analysis (se f.eks. Dahler-Larsen og Krogstrup 2009, Krogstrup 2006, Dybdahl 2009 og EVA 2009a). Langt sjældnere, i hvert fald uden for fagøkonomiske kredse, har der været en seriøs diskussion af alternativet til disse tilgange. Alternativet kan kaldes den kontrafaktiske tilgang til kausalitet, på engelsk *the Neyman-Rubin counterfactual framework* efter statistikerne Neyman og Rubin. Selvom denne tilgang er udviklet af statistikere, har den desuden rod i flere forskellige discipliner, f.eks. økonomi og psykologi, og på det seneste har den også gjort sit indtog i sociologien (Guo og Fraser 2010, Winship og Morgan 2007).

Den kontrafaktiske tilgang

Ideen med den kontrafaktiske tilgang er at vi ikke kan tale om effekten af en indsats uden at vi har et kvalificeret bud på hvad der ville være sket hvis vi ikke havde gennemført indsatsen. Dette forhold – hvad der ville være sket hvis vi ikke havde gennemført indsatsen – kaldes den kontrafaktiske situation. For at kunne fastslå hvad effekten af en mentorordning er på en række udvalgte skoler, må vi vide hvilket frafald der ville have været på de samme skoler uden en mentorordning. For evaluatorene der har arbejdet med effektevaluering, giver dette intuitivt god mening. Vi ved godt at effekten af et jobkabelsesprogram ikke kan gøres direkte op som andelen af klienter der er i arbejde efter en given periode. Evaluerer vi f.eks. en privat aktør i beskæftigelsesindsatsen med ansvar for 100 arbejdsløse akademikere, kan vi ikke opgøre effekten ved alene at se på hvor mange af dem der er i arbejde efter en periode på f.eks. 12 måneder. Vi ved at en del af disse akademikere ville have fundet arbejde uanset om de havde deltaget i den private aktørs jobkabelsesprogram eller ej. Måske har programmet ligefrem forhindret nogle i at få job ved at optage deres tid med andre aktiviteter end at søge arbejde. I dette tilfælde ville effekten af jobkabelsesprogrammet være negativ, også selvom den private aktør på papiret ville kunne fremvise nogle umiddelbart imponerende resultater i forhold til hvor mange akademikere der var i arbejde efter 12 måneder. For at kunne udtale os om effekten af jobkabelsesprogrammet bliver vi

nødt til at have en forestilling om den kontrafaktiske situation, i dette tilfælde hvor mange af de arbejdsløse akademikere der havde fundet et job uden den private aktørs hjælp.

Når man evaluerer effekten af en indsats ud fra den kontrafaktiske tilgang til kausalitet, består den metodiske indsats i at estimere hvordan den kontrafaktiske situation ville have set ud. Optimalt ville vi gerne evaluere effekten af en indsats ved at se på det samme individ i to forskellige situationer. Det er imidlertid ikke muligt da man ikke på samme tid både kan have en mentor og ikke have en mentor eller på samme tid både deltage i et jobskabelsesprojekt og ikke deltage i et jobskabelsesprojekt. Derfor er den kontrafaktiske situation en teoretisk konstruktion som ikke kan observeres, men kun estimeres. I stedet for at sammenligne udfaldet for samme deltagere sammenligner vi udfaldet for sammenlignelige grupper af deltagere. Med sammenlignelige grupper menes grupper der er sammenlignelige i statistisk forstand.

Lodtrækningsforsøget

Det bedste design i forhold til at estimere den kontrafaktiske situation er også det mest simple. Designet kaldes ”lodtrækningsforsøget” eller på engelsk *the randomized controlled trial*. Har vi mulighed for at placere 200 unge i to forskellige grupper ud fra en tilfældig lodtrækning hvor den ene gruppe (indsatsgruppen eller interventionsgruppen) modtager indsatsen og den anden gruppe (kontrolgruppen) ikke gør, ved vi, såfremt randomiseringen er foretaget korrekt og alt i forbindelse med forsøget forløber som planlagt, ikke at de to grupper af ens, for det er der ikke nogen grupper der er, men at forskellene mellem dem er tilfældige. Den eneste systematiske forskel mellem de to grupper er at den ene gruppe har modtaget indsatsen og den anden ikke har. Når vi bagefter analyserer forskellene mellem grupperne, kan vi ved hjælp af en simpel signifikanstest analysere om de forskelle vi iagttager mellem de to grupper, er så store at vi ved et givent signifikansniveau tør tro på at de ikke er opstået på grund af tilfældige forskelle mellem de to grupper, men at de skyldes indsatsen.

At lodtrækningsforsøget bygger på en tilfældig fordeling af deltagerne, er netop designets store force i metodisk henseende. Frem for at skulle kontrollere for betydningen af andre variable, kan vi lade randomiseringen gøre arbejdet for os. I mange tilfælde er det imidlertid ikke praktisk muligt, eller etisk forsvarligt, at gennemføre et lodtrækningsforsøg for at evaluere effekten af indsats. Og hvad gør man så? Her er matching en mulig vej.

Matching

Ideen i matching er at matche hvert enkelt individ der deltager i en indsats, med et eller flere individer der ligner det første individ så meget som muligt, men som ikke har deltaget i indsatsen. På den måde opbygges der ad statistisk vej en kontrolgruppe der ligner den anden så godt at den er et godt udtryk for den kontrafaktiske situation.

Mere formelt kan man sige at hvis vi lader y_1 betegne outcome med indsatsen og lader y_0 betegne outcome uden indsatsen, er det $E(y_1 - y_0)$ vi ønsker at estimere. Her betegner E forventningsoperatoren eller ”treatmenteffekten”. Denne angiver altså den *gennemsnitlige* effekt af indsatsen i den gruppe af personer der har modtaget indsatsen. Problemet er, som det også blev nævnt ovenfor, at vi for hvert enkelt individ kun observerer enten y_1 eller y_0 og ikke begge.¹

¹ En grundig matematisk indføring i teorien bag matchingmetoder kan findes i Rosenbaum og Rubins oprindelige artikel eller i Wooldridge (2002), kapitel 18.

Når man foretager såkaldt *simpel matching*, identificerer man faktorer der både har betydning for deltagelse i indsatsen og outcome, og foretager matching ud fra dem. Har køn en betydning for deltagelse i indsatsen outcome, matcher man hver kvinde i indsatsgruppen med en kvinde i kontrolgruppen; har etnicitet en betydning, matcher man hver efterkommer af en indvandrer med en efterkommer osv.

Der behøver blot at være et par baggrundsfaktorer som man ønsker at matche i forhold til, før det bliver vanskeligt at foretage simpel matching. Det kan være vanskeligt at matche hvert individ i indsatsgruppen med et individ i kontrolgruppen når en række kriterier skal opfyldes på én gang. Man kan havne i situationer hvor man f.eks. leder efter et individ der er kvinde, er i aldersgruppen 15-16 år, studerer på en bestemt uddannelse, hvis forældre er ufaglærte, og som tilhører en bestemt etnisk minoritet. Dette vil betyde at mange individer i indsatsgruppen ikke kan matches med et individ i kontrolgruppen, og man bliver dermed tvunget til det metodisk beklagelige at måtte udelade dem fra evalueringen. Løsningen på dette problem er at matche alle på et udtryk der samler indflydelsen fra de forskellige baggrundsfaktorer der har betydning.

Statistikerne Rosenbaum og Rubin påviste i 1983 at vi kan anvende den såkaldte *propensity score* for en bestemt gruppe individer som et samlet udtryk for baggrundsfaktorerne. Propensity scoren udregnes som sandsynligheden for at deltage i en indsats givet individets baggrundskaraktistika. Hvis det lykkes os at estimere propensity scoren korrekt, kan man estimere treatmenteffekten simpelt som gennemsnittet af outcome blandt individer der har modtaget indsatsen, minus gennemsnittet af outcome blandt individer der ikke har modtaget indsatsen, men som har samme propensity score. I praksis kan det være svært eller umuligt at finde individer med nøjagtigt samme propensity score, hvorfor der oftest anvendes et mindre strengt kriterium, nemlig at propensity scores ikke må være for forskellige.

Propensity score matching i praksis

I praksis kan propensity score matching f.eks. gennemføres ved at følge nedenstående fremgangsmåde der let kan implementeres i STATA gennem proceduren `psmatch2`:

1. Identificer individerne i indsatsgruppen
2. Identificer de individer der potentielt set kan indgå i en kontrolgruppe
3. Estimer propensity scoren ved hjælp af f.eks. en probitregression eller en logistisk regression for samtlige individer der er identificeret under 1 og 2.
4. Udvælg en kontrolgruppe fra gruppen af individer der potentielt set kan indgå i en kontrolgruppe ved at udvælge de individer der har en propensity score der matcher individerne i treatmentgruppen bedst muligt.
5. Estimer effekten ud fra en sammenligning af et simpelt gennemsnit de to grupper imellem. Om denne effekt er signifikant, afgøres af en simpel t-test.

Når treatmentgruppen skal identificeres, er det vigtigt at gøre sig klart præcis hvad der menes med at et individ har deltaget i en indsats. F.eks. kan der tænkes eksempler på at et bestemt individ har deltaget i mere end én indsats. Skal dette individ indgå i analysen eller ej? Svaret afhænger af om man vil være sikker på at måle den rene effekt af en enkelt indsats, eller om man vil være tilfreds med at kunne måle en samlet effekt af flere indsatser.

Når man skal identificere den mulige kontrolgruppe, må man gøre sig klart hvilken målgruppen indsatsen er rettet mod. Hvis man f.eks. vil evaluere en indsats der er målrettet mod

gymnasieelever, skal man formodentlig begrænse den mulige kontrolgruppe til at indeholde alle gymnasieelever. Men hvis deltagerne i indsatsen kun er kvindelige gymnasieelever, skal de mandlige elever nok tages ud af den mulige kontrolgruppe.

Estimation af propensity score foregår ved at man estimerer en sandsynlighed for deltagelse i indsatsen givet de baggrundskarakteristika der foreligger oplysninger om. Den præcise estimationsmetode kan variere, men som oftest anvendes probitestimation eller logistisk estimation. Som resultat af denne estimation fås en estimeret sandsynlighed – propensity score – for deltagelse i indsatsen der ligger mellem 0 og 1. Denne sandsynlighed foreligger for såvel deltagere som ikke-deltagere.

Udvælgelsen af kontrolgruppen sker ved for hvert individ i treatmentgruppen at udvælge det individ i den mulige kontrolgruppe hvis propensity score ligger tættest på. Dette kan betyde at ét individ kan komme til at indgå i kontrolgruppen to gange.

Som påvist af Rosenbaum og Rubin er det herefter let at estimere treatmenteffekten idet vi blot tager gennemsnittet for treatmentgruppen og sammenligner med gennemsnittet for kontrolgruppen. En simpel t-test kan vise om disse to størrelser er statistisk signifikant forskellige.

I praksis anvendes der ofte en kombination af simpel matching og propensity score matching. Da CEBR f.eks. evaluerede innovationsindsatsen over for små og mellemstore virksomheder, var det en betingelse at virksomheden i kontrolgruppen skulle være fra samme branche som virksomheden i treatmentgruppen. Dermed undgik man at sammenligne virksomheder med vidt forskellig produktion som blot tilfældigvis havde samme propensity score.

Den praktiske gennemførelse af matching illustreres i det følgende ved hjælp af to eksempler ².

Effekten af AA-møder på alkoholmisbrug

Et hospital i USA havde kontakt til 218 krigsveteraner med alkoholproblemer. Veteranerne fik tilbud om at deltage i et afvænningsprogram afholdt af Anonyme Alkoholikere (AA). For de veteraner der havde lyst til at deltage, ville hospitalet afholde udgiften. Af de 218 tog nogle af veteranerne imod tilbuddet og deltog i møderne, andre gjorde ikke.

Når vi ønsker at foretage en evaluering af effekten af at deltage i afvænningsprogrammet, kan vi ikke bare sammenligne de veteraner der deltog i AA-programmet, med gruppen af veteraner der ikke deltog. Hele filosofien i AA bygger på at en alkoholmisbruger selv skal vælge at deltage i programmet og at engagere sig i behandlingen. De veteraner der melder sig til at deltage i programmet, må derfor formodes at være mere motiverede for at bryde med deres alkoholmisbrug end andre veteraner. Med andre ord er gruppen af veteraner der ikke deltager i programmet, som helhed et dårligt udtryk for den kontrafaktiske situation.

For at opnå et ikke-biased udtryk for den kontrafaktiske situation anvender vi matching. Der var en række demografiske forskelle mellem den gruppe der valgte at deltage i AA-programmet, og de øvrige veteraner. F.eks. var andelen af kvinder der deltog i AA-møderne, større end i gruppen af de øvrige veteraner. Vi kunne derfor godt foretage en simpel matching i forhold til køn og andre faktorer. Men som nævnt er det, når man ønsker at matche i forhold til en række variable, vanskeligt

² En redegørelse for forskelle og ligheder mellem propensity score matching og beslægtede teknikker som Heckmans sample selection model og matching estimators findes i Guo og Fraser 2010.

at finde et præcist match til hver deltager i indsatsgruppen, og vi kan derfor blive tvunget til at opgive nogle af observationerne i indsatsgruppen. I stedet forsøger vi at matche deltagere i interventionsgruppen med veteraner der ikke deltog i AA-møderne, men som egentlig havde samme sandsynlighed for at gøre det. Af en eller anden – antaget tilfældig – grund gjorde de det bare ikke.

Sandsynligheden for at ville deltage i møderne er forskellig fra veteran til veteran. For nogle vedkommende er det meget sandsynligt at de vil tage imod tilbuddet, for andres vedkommende er det meget usandsynligt. Vi udtrykker sandsynligheden som et tal mellem 0 og 1. Sandsynligheden hænger sammen med en række baggrundsfaktorer. Hospitalet fandt ud af at følgende faktorer havde en betydning for om veteranerne ønskede at deltage i programmet:

- Om man er mand eller kvinde (kvinder havde større sandsynlighed for at deltage end mænd)
- Hvor alvorligt man bedømmer sit alkoholproblem på en skala fra 1 til 10 (jo alvorligere man bedømmer sit problem, desto mere sandsynligt er det at man deltager i indsatsen)
- Hvor tilbøjelig man er til at løse problemer ved at søge hjælp fra andre på en skala fra 1 til 10 (jo mere tilbøjelig man er, desto højere er sandsynligheden for at man deltager i indsatsen).

Disse data samler vi ind for alle 218 deltagere. For alle deltagere beregnes herefter propensity scoren for at deltage i programmet, uanset om de rent faktisk deltog i programmet eller ej.

Derefter matcher vi hvert individ i gruppen af deltagere i AA-møderne med et individ fra gruppen af veteraner der ikke deltog i møderne, men som havde samme sandsynlighed for at deltage i møderne.

På denne måde får vi skabt en kontrolgruppe der er et godt udtryk for den kontrafaktiske situation. Herefter forholder vi os til forskellen mellem interventionsgruppe og kontrolgruppe i forhold til vores outcome. Målt på et indeks for alvorligheden af alkoholproblemer der går fra 1 til 30, var der en signifikant forskel på over 6 point mellem interventionsgruppen og den matchede kontrolgruppe. Deltagelse i AA-møderne var altså et effektivt middel mod alkoholproblemer (eksemplet er inspireret af Rossi et al. 1999, 328-329).

Måling af effekten af aktivering i forhold til at skaffe ledige i arbejde

Et andet eksempel på brug af propensity score matching er en evaluering af aktiveringsindsatsen over for forsikrede ledige (CEBR, 2009). I Danmark har ledige ret og pligt til aktivering. Det betyder at man skal tilbydes et aktiveringstilbud når man har været ledig i en vis periode, og at man har pligt til at gennemføre en sådan aktivering hvis man vil bevare retten til at modtage dagpenge. Imidlertid er det omdiskuteret om denne aktiveringsindsats virker. Får aktiveringsperioderne flere i beskæftigelse? Eller fastholder de måske ligefrem personer i ledighed?

I denne situation er det forholdsvis let at identificere treatmentgruppen som alle de individer der har været i et bestemt type aktiveringsforløb inden for en bestemt periode. Med den mulige kontrolgruppe forholder det sig lidt anderledes. Hvilke ledige skal indgå? Skal det være alle, eller skal det kun være nogle med bestemte karakteristika? I den givne evaluering blev det besluttet at tage udgangspunkt i alle som var ledige i første uge af 2002. De som efterfølgende blev aktiveret, udgjorde treatmentgruppen, mens de der ikke blev aktiveret, udgjorde kontrolgruppen.

Ved estimering af propensity score afsløredes følgende karakteristika blandt individerne i treatmentgruppen, dvs. de individer som havde deltaget i jobtræning i en privat virksomhed:

- Den gennemsnitlige ledighedsanciennitet var næsten 30 uger længere i treatmentgruppen. Dette skyldes at aktivering først tilbydes efter en vis periodes ledighed.
- Blandt de aktiverede var kun 44 % kvinder, mens 54 % blandt de ikke-aktiverede var kvinder.
- Blandt de aktiverede var 43 % ufaglærte, mens 38 % blandt de ikke-aktiverede var ufaglærte.

Disse forskelle, som det kunne lade sig gøre at rette op på ved hjælp af propensity score matching, viser at kontrolgruppen og treatmentgruppen adskilte sig på en række områder der er relevante for muligheden for at komme i beskæftigelse. Den kontrafaktiske situation er anderledes for en gruppe med 43 % ufaglærte end for en gruppe med 38 % ufaglærte, hvorfor det er vigtigt at kontrollere for denne forskel.

Resultatet af målingen blev at jobtræning i en privat virksomhed i første omgang viste sig at have en negativ effekt på beskæftigelsen. Godt et år efter evalueringens start var der imidlertid en positiv effekt. Andelen af beskæftigede var højere i interventionsgruppen end i kontrolgruppen. Analysen viser dermed at aktivering i private virksomheder på længere sigt er effektivt i forhold til at skaffe ledige i arbejde.

Metoden stiller krav til implementering og data

En nylig afsluttet evaluering hvor propensity score matching er anvendt som metode, viser at en meningsfuld anvendelse af metoden forudsætter at en række krav er opfyldt (EVA 2009b). For det første skal indsatsen være veldefineret og velafgrænset. Dette lyder tilsyneladende simpelt, men i praksis er der på mange områder inden for offentlig forvaltning og service stor variation i indholdet af indsatser der betegnes på samme måde. I den evaluering der blev omtalt i artiklens begyndelse, ønskede man f.eks. at undersøge om indførelsen af en mentorordning på erhvervsuddannelserne nedbringer elevfrafaldet på grundforløbet. Det viste sig imidlertid at det var meget forskelligt fra skole til skole hvad det indebar at have en mentorordning. Nogle skolars mentorordning bestod i elev til elev-mentorer, på andre skoler havde en lærer afsat nogle timer i sit skema til at arbejde med fastholdelse af elever. På andre skoler havde man en socialpædagog ansat på fuld tid, og på atter andre havde en pensioneret håndværker eller lignende en eller anden form for kontakt med frafaldstruede elever. Det varierede også om mentoren var tilknyttet en række faste elever, eller om mentorordningen var et tilbud til de af skolens elever der følte et behov. Ligeledes varierede intensiteten af ordningen; på nogle skoler havde mentorerne mange arbejdstimer, på andre skoler få. Alle disse forhold gav slør og upræcise estimater i forhold til at vurdere mentorordningernes effekt generelt.

En anden erfaring fra denne undersøgelse er at man for at estimere effekten af tiltag der retter sig mod bestemte personer, er nødt til at have oplysninger om hvem disse personer er (f.eks. deres cpr-nummer). Det giver for stor usikkerhed at ville undersøge effekten af indførelsen af en mentorordning på skolens aggregerede frafaldsniveau.

En tredje erfaring er at metoden kan være vanskelig at bruge hvis der er tale om et område hvor der bliver igangsat mange indsatser. Vanskeligheden hænger sammen med konstruktionen af både indsatsgruppe og kontrolgruppe. Hvad angår indsatsgruppen, er det vanskeligt at afgøre præcis hvad

der virker, hvis der på skolen f.eks. er igangsat både en mentorordning og et tilbud om psykologhjælp, og hvis der samtidig er afsat penge til at forbedre det sociale miljø. Kombinationen af indsatser vil derudover variere fra skole til skole. Dette vanskeliggør også konstruktionen af en kontrolgruppe der er ”ren” og ikke påvirket af effekten af andre indsatser.

De nævnte forhold vedrører i bund og grund ikke metodens validitet, men måden vi implementerer politikker og policies på i Danmark. Det fører til en række spørgsmål om sammenhæng mellem dette og et stærkt ønske fra politisk hold om at få viden om effekter af forskellige indsatser.

Metodisk diskussion

Matching kan kritiseres på de samme punkter som andre kvasiekperimentelle metoder. Først og fremmest er det som gennemgået ovenfor en vigtig forudsætning at deltagelse i indsatsen og outcome er uafhængige for individer med samme propensity score. Er dette ikke tilfældet, er de estimerer der fremkommer ved brug af metoden, biased. I praksis vil dette ofte skyldes at ikke alle variable der er af betydning for selektionen til treatment, er observeret og medtaget i selektionsmodellen. Hvis det i eksemplet med aktiveringsindsatsen f.eks. er sådan at personer med et bestemt ”drive” er mere tilbøjelige til at komme i privat jobtræning, og hvis dette drive også har betydning for deres udbytte af indsatsen, er betingelsen om uafhængighed ikke opfyldt, idet vi i analysen ikke har undersøgt for dette drive. Kun hvis netop de variable vi har observeret, kan forklare et sådant drive, f.eks. hvis personernes drive er afhængigt af erhvervs erfaring, uddannelse eller lignende, vil vi alligevel kunne sige at betingelsen er opfyldt.

Eksemplet viser at validiteten af en matchingestimator i høj grad afhænger af om vi kan observere de variable der er relevante for selektionen og udfaldet. Hvis der er uobserverede variable der må forventes at spille en afgørende rolle for såvel selektion som outcome, vil to individer der ser ud til at have samme sandsynlighed for at deltage i en indsats, i realiteten ikke have det. Og dermed vil man stå med et forkert estimat af den kontrafaktiske situation.

For at undersøge betydningen af hvilke faktorer man matcher i forhold til, har Shadish et al. (2008) gennemført en undersøgelse hvor de har estimeret effekten af en indsats både ved hjælp af lodtrækningsforsøg og ved hjælp af kvasiekperimentelle metoder, herunder propensity score matching. Det viser sig at propensity score matching er i stand til at levere resultater der er næsten (96 %) lige så præcise som et lodtrækningsforsøg hvis man er i stand til at identificere og modellere de faktorer der har betydning for selektionen. Matcher man derimod i forhold til såkaldte *factors of convenience*, altså forhåndenværende baggrundsdata som køn, alder osv. som i mindre grad er korreleret med selektion og outcome, vil resultaterne af matchingen være upræcise og biased. Gør man dette, er der stor risiko for at nå den forkerte konklusion om effektiviteten af en indsats og dermed gøre mere skade end gavn ved at anbefale videreførelse af ineffektive indsatser og nedlæggelse af effektive indsatser.

Problemet med at anvende en kvasiekperimentel metode som matching er at man normalt har begrænsede muligheder for at undersøge om ens modellering af selektionen er god, eller om der er vigtige uobserverede variable man ikke har taget højde for.

Til trods for at propensity score matching også er forbundet med udfordringer, er der dog ingen tvivl om at der er mange tilfælde hvor metoden kan anvendes i evalueringssøjmed. I Danmark har vi særligt gode muligheder for at bruge metoden da vi gennem Danmarks Statistiks registre kan få

adgang til detaljerede oplysninger om alle individer og virksomheder i landet. Dermed har vi bedre muligheder for at skabe kontrolgrupper gennem matching end man har i de fleste andre lande.

Litteratur:

CEBR (2009): *Analyser af effekten af aktivering og voksen- og efteruddannelse for forsikrede ledige*. København.

http://www.cebr.dk/upload/analyser_af_aktivering_og_veu_for_forsikrede_ledige.pdf

Dahler-Larsen, Peter og Krogstrup, Hanne Katrine (2009): *Nye Veje i Evaluering*, Viborg: Akademia.

Dybdahl, Line (2009): "Contribution Analysis – et alternativ til klassiske effektdesigns?", *Evalueringssnyt*, 24, pp. 12 -15.

EVA (2009a): *Viden der forandrer – Virkningsevaluering af læsevejlederen som fagligt fyrtårn*. København.

<http://www.eva.dk/projekter/2008/virkningsevaluering-af-laesevejledning-i-en-kommune/projektprodukter/viden-der-forandrer.-virkningsevaluering-af-laesevejlederen-som-fagligt-fyrtaarn>

EVA (2009b): *Frafald på grundforløbet på de merkantile erhvervsuddannelser*. København.

<http://www.eva.dk/projekter/2008/effektevaluering-af-fracald-paa-merkantile-grundforloeb/rapport/fracaldet-paa-grundforloebet-paa-de-merkantile-erhvervsuddannelser>

Guo, Shenyang og Fraser, Mark W. (2010): *Propensity Score Analysis. Statistical Methods and Applications*, London: Sage.

Krogstrup, Hanne Kathrine (2006): *Evalueringssmodeller*, 2. udgave. Århus: Academia.

Munk, Martin (2008): Metoder til at måle kausale effekter af socialpolitiske indsatser, *Dansk Sociologi*, 19(1), pp. 55-73.

Rosenbaum, P.R. og Rubin D.B. (1983): The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika* 70(1), pp. 41-55.

Rossi, Peter et al. (1999): *Evaluation. A systematic approach*, 6th Edition. London: Sage.

Shadish, William, Clark, M.H. og Steiner, Peter (2008): "Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random to Nonrandom Assignment", *Journal of the American Statistical Association*, 103. pp. 1334 - 1343.

Winship, Christopher og Morgan, Stephen L. (2007): *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, Cambridge: CUP.

Wooldridge, J.M. (2002): *Econometric Analysis of Cross Section and Panel Data*, Cambridge, Massachusetts: MIT Press.