Metaanalyse

Geir Smedslund

Seksjon for spesialisthelsetjenesten, Nasjonalt kunnskapssenter for helsetjenesten

E-post: geir.smedslund@kunnskapssenteret.no Telefon: 92 43 01 24

SAMMENDRAG

Metaanalyse er en kvantitativ metode for å oppsummere resultatene av flere enkeltstudier. I en metaanalyse forsøker man å tallfeste behandlingseffekten, og man gir store studier større vekt enn små studier. En mye brukt metode for å vekte er invers variansmetoden. Dersom alle studiene har målt resultatene på samme måte kan resultatene brukes direkte i metaanalysen, men dersom det samme utfallet er målt på ulike måter, må man bruke standardiserte effektstørrelser hvor alle resultatene er omregnet til en felles skala. Dersom man tror at effekten av behandlingen vil være lik for alle, bortsett fra tilfeldige variasjoner, benytter man en fixed-effect modell. Tror man derimot at det vil være systematiske forskjeller i effekt når behandlingen gis i ulike kontekster, legges dette inn i en såkalt random-effects modell. Metaanalyser blir ofte fremstilt grafisk i form av forest plots. Hver linje representerer da én studie, med effektestimatet markert som et punkt, mens ytterpunktene av linjen representerer konfidensintervallet. Metaanalysen blir fremstilt som en diamant hvor bredden viser usikkerheten i estimatet. Dersom resultatene fra alle studiene trekker i samme retning er metaanalysen "homogen". Men dersom studiene spriker når det gjelder effektstørrelse og retning på effekt, er det "heterogenitet". Styrken ved metaanalyse er at den kan sammenfatte en stor mengde informasjon i ett tall. Samtidig er dette også svakheten ved metoden. Et enkelt tall kan ikke beskrive variasjonen på tvers av flere studier.

Smedslund G. **Meta-analysis.** *Nor J Epidemiol* 2013; **23** (2): 147-149.

ENGLISH SUMMARY

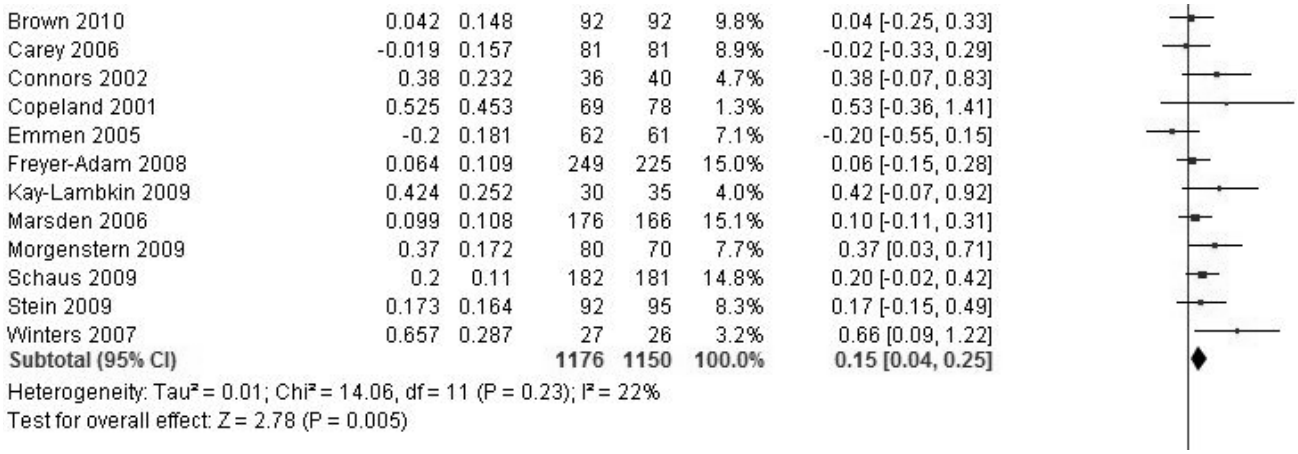
Meta-analysis is a quantitative method for summarizing single studies. In a meta-analysis, one tries to quantify the treatment effect, assigning more weight to large studies than to small studies. A much used method for weighting is the inverse variance method. If all studies have measured the results in the same way, the results can be used directly in the meta-analysis, but if the same outcome is measured in different ways across different studies, one has to use a standardized effect size where results are converted to a common scale. If it is believed that the effect is consistent across various populations and settings, one can employ a fixed-effect model. If systematic differences in effect can be expected, a random-effects model is used. Meta-analyses are often depicted as forest plots. Each line represents one study where the effect estimate is marked as a point on a line, with each end of the line representing the confidence interval around it. The meta-analysis is shown as a diamond where the width illustrates the uncertainty around the estimate. If all study results point in the same direction, the meta-analysis is considered "homogeneous". But if the studies vary in their effect size and direction, the findings are "heterogeneous". The strength of meta-analysis is that it can be used to summarize a large body of information in one number. This is also its limitation. One number cannot describe the variation that exists across different studies.

Ordet metaanalyse ble først lansert av Glass (1) selv om metoden ikke var ny. Metaanalyse betyr "analyse av analyse" og henspiller på at man analyserer primærstudier som igjen har analysert for eksempel pasienter. Som eksempel skal jeg i dette kapitlet bruke en metaanalyse fra Cochrane-oversikten *Motivational Interviewing for Substance Abuse (2)*.

ULIK VEKTING AV ULIKE STUDIER

I denne Cochrane-oversikten inkluderte vi 12 studier der forskerne hadde forsøkt å evaluere effekten av motiverende intervju (MI) på rusmiddelbruk, målt 6-12 måneder etter det motiverende intervjuet. Av de 12 studiene var det bare to studier som rapporterte en statistisk signifikant effekt av MI ($p<0,05$). I hver av de resterende 10 studiene var det ingen signifikant forskjell i rusmiddelbruk mellom de som hadde blitt randomisert til MI og de som var i kontrollgruppen. Dersom man teller opp resultatene på denne måten, kalles det "vote counting" og det anbefales ikke. Hvis man bare teller opp hvor mange studier som fant effekt og hvor mange som ikke fant effekt, tar man ikke hensyn til hvor stor effekten er. Man tar heller ikke hensyn til at noen studier er større enn andre. I en metaanalyse forsøker man å tallfeste hvor stor effekten er, og man gir store studier større vekt enn små studier.

I to av studiene var det faktisk litt mindre rusmiddelbruk i kontrollgruppen selv om forskjellen ikke var signifikant. I tabell 1 vises resultatene fra hver enkelt studie. Studiene benyttet litt ulike måter å måle effekt av MI. For å kunne sammenlikne resultatene ble derfor



Figur 1. Den fullstendige metaanalysen fra Cochrane-oversikten *Motivational Interviewing for Substance Abuse (2).* Gjengitt med tillatelse fra John Wiley & Sons Ltd.

Tabell 1. Studiene som inngikk i metaanalysen (2).

Study	N Intervention	N Control	Standardized Mean Effect Size (SE)
Brown 2010	92	92	0.042 (0.148)
Carey 2006	81	81	–0.019 (0.157)
Connors 2002	36	40	0.38 (0.232)
Copeland 2001	69	78	0.525 (0.453)
Emmen 2005	62	61	–0.2 (0.181)
Freyer-Adam 2008	249	225	0.064 (0.109)
Kay-Lambkin 2009	30	35	0.424 (0.252)
Marsden 2006	176	166	0.099 (0.108)
Morgenstern 2009	80	70	0.37 (0.172)
Schaus 2009	182	181	0.2 (0.11)
Stein 2009	92	95	0.173 (0.164)
Winters 2007	27	26	0.657 (0.287)

effektene målt som en "standardisert forskjell". I studien til Brown 2010 er denne på 0,042, som betyr at forskjellen i rusmiddelbruk mellom MI-gruppen og kontrollgruppen var omtrent 0,04 standardavvik. Tallene i parentes (0,148 hos Brown 2010) er standardfeilen som angir hvor presis effektstørrelsen er. En stor standardfeil betyr at effekten er upresis. Standardfeilen henger sammen med hvor mange som deltok i studiene. Vi har tre store studier (Marsden 2006, Schaus 2009 og Freyer-Adam 2008) med små standardfeil (henholdsvis 0,108, 0,110 og 0,109), mens en av de minste studiene (Copeland 2001) har den største standardfeilen (0,453). Standardfeilen blir mindre når det er mange deltakere i en studie, men i noen studier er det større forskjell mellom deltakerne enn i andre studier. I en studie kan det for eksempel være slik at alle har omtrent like stor nytte av behandlingen, mens i andre studier kan det være noen som har stor bedring og andre kanskje heller har blitt verre enn bedre.

I eksemplet som brukes i dette kapitlet måtte man gå en omvei ved å bruke standardisert forskjell, men ofte kan man bruke resultatene fra primærstudiene direkte i metaanalysen. Dersom alle studiene for eksempel hadde antall drinker per dag som utfall, ville det ikke vært noe i veien for å bruke dette direkte. Det ville også gjort metaanalysen lettere å tolke.

Ofte bruker man den såkalte "invers varians metoden" for å vekte studiene. Formelen er 1/SE (1 delt på standardfeilen). Den lille studien Copeland 2001 vil for eksempel få vektingen 1/0,453 = 2,2. Den store studien Marsden 2006 vil få en vekt på 1/0,108 = 9,3. I Figur 1 ser vi vektingen uttrykt som prosent slik at summen av vektene blir 100 prosent. Copeland 2001 får her bare 1,3 prosent av vektingen, mens Marsden 2006 får 15,1 prosent.

"FIXED EFFECT" ELLER "RANDOM EFFECTS"?

Det er viktig å bestemme seg for hva man skal forsøke å måle med en metaanalyse, og det er i hovedsak to tilnærminger: Den første går ut på å regne ut *effekten* av behandlingen – en "fixed effect". Det antas da at behandlingen har samme effekt i alle studier, bortsett fra tilfeldige variasjoner, og metaanalysen gjøres med en "fixed-effect modell". Den andre tilnærmingen legger *ikke* til grunn at effekten er den samme, men at den for eksempel vil variere dersom den gis i ulike kontekster, til ulike pasientgrupper osv. I så fall vil en forvente forskjeller i resultater fra studie til studie, og det vil være *gjennomsnittseffekten* på tvers av studiene en er ute etter å måle. Da brukes en "random-effects modell".

GRAFISK FRAMSTILLING

I Figur 1 vises et såkalt forest plot helt til høyre. Den vertikale linjen i Figur 1 representerer ingen forskjell mellom de som fikk MI og de som var i kontrollgruppen. Alle punkter til høyre for linjen indikerer at MI gruppen gjorde det best, mens punkter til venstre (Carey 2006 og Emmen 2005) betegner at kontrollgruppen gjorde det best. De horisontale linjene viser konfidensintervallene rundt hvert av effektestimatene.

Den lille og upresise studien Copeland 2001 har et bredt konfidensintervall, fra -0,36 til 1,41. Fordi konfi-

densintervallet krysser den vertikale linjen, ser vi øyeblikkelig at resultatet for denne studien ikke er signifikant. Vi ser også at konfidensintervallet til Marsden 2006 er veldig smalt, noe som betyr at resultatet her er ganske presist. Men siden det går fra -0,11 til 0,31 er dette resultatet heller ikke signifikant. Derimot er hele konfidensintervallet for Winters 2007 på høyre side av den vertikale linjen – resultatet er signifikant i favør av MI.

Den nederste linjen i Figur 1 er resultatene fra selve metaanalysen. Her oppsummeres alle studiene: I alt var det 1175 personer som fikk MI mens det var 1150 i kontrollgruppene. Den standardiserte forskjellen i metaanalysen er 0,15 med konfidensintervall fra 0,04 til 0,25. Dette ser vi helt til høyre som en (sort) diamant. Hele diamanten er på høyre side av midtlinjen, og dette betyr at metaanalysen er signifikant i favør av MI. Signifikansen ser vi også som "Test for overall effekt" med en p-verdi på 0,005.

Dersom alle studiene trekker i samme retning, sier vi at metaanalysen er homogen. Men dersom studiene spriker når det gjelder effektstørrelse og retning på effekt, sier vi at det er stor heterogenitet. Nederst i figur 1 står det noen data om akkurat dette, blant annet resultatet av en Chi-kvadrat test for om alle studiene har fått samme resultat (bortsett fra forskjeller som skyldes tilfeldige forskjeller mellom deltakerne). P-verdien for testen er 0,23, noe som tyder på liten grad av heterogenitet. Vi har også noe som kalles I-kvadrat (I^2). Denne er på 22 prosent. Kort forklart betyr dette at 22 prosent av den totale variansen i metaanalysen er varians *mellom* studiene. De gjenstående 78 prosentene er varians mellom deltakerne *innenfor* de enkelte studiene.

Hvor stor er effekten av MI? Bruk av standardiserte forskjeller som uttrykk for effektstørrelse kan gjøre det vanskelig å forstå hva resultatene betyr i praksis. En hyppig brukt tommelfingerregel for fortolkning av standardiserte forskjeller skriver seg fra en bok av Jacob Cohen (3). Han skriver at en standardisert forskjell på 0,2 er å regne som en liten effekt, mens 0,5 er en middels effekt, og 0,8 kan ses på som en stor effekt. Ser vi på studiene i vår metaanalyse ser vi for eksempel at Copeland 2001 har funnet en middels effekt, mens effekten i Winters 2007 er litt større enn middels. Schaus 2009 fant en liten effekt, mens Emmen 2005 fant en liten effekt i favør av kontrollgruppen. Metaanalysen som helhet har funnet en effekt som er litt mindre enn en typisk liten effekt, om vi legger Cohen sine kriterier til grunn.

STYRKER OG SVAKHETER VED METAANALYSER

Den mest opplagte styrken ved en metaanalyse er at man ved å slå sammen resultater fra flere studier kan regne ut en effektstørrelse som er mer presis enn det man får fra én studie. Jo flere studier som legges inn, desto mer presis blir effektstørrelsen. Den eksterne validiteten blir også større ved en homogen metaanalyse enn ved en enkelt studie: Én studie har alltid én type pasienter som får én type behandling i én kontekst. Metaanalysen, derimot inneholder informasjon om hvordan effekten varierer fra én type pasienter til en annen type, fra én type behandling til en annen type, osv.

Den største svakheten ved metaanalyser er nettopp det samme som styrken; at man oppsummerer store datamengder i form av ett enkelt tall. Det er et problem at resultatene fra en studie bare angir hvordan det går med gjennomsnittspasienten, og derfor ikke forteller oss hvordan virkningen vil bli for en spesifikk pasient. I en metaanalyse blir dette problemet forsterket: resultatet angir virkningen som et gjennomsnitt av gjennomsnittspasientene. En annen svakhet ved metaanalyser er at dersom det er stor heterogenitet gir det ingen mening å regne ut et gjennomsnitt. Dette problemet kan man noen ganger løse ved å dele opp metaanalysen. I vårt eksempel er det to studier (Carey 2006 og Emmen 2005) som går i motsatt retning av de andre. Hvis man fant at det var spesielle kjennetegn ved disse to studiene som skilte dem fra de andre, kunne man ta dem ut og lage en egen metaanalyse med bare disse to studiene. Det ville føre til at heterogeniteten blant de gjenværende studiene ble mindre. Et annet problem ved metaanalyser er at dersom studiene som inngår i analysen har metodiske svakheter som innebærer at resultatene ikke er til å stole på, kan man ikke stole på resultatene fra metaanalysen heller.

Eksemplet vårt har vist at metaanalyse kan fremskaffe ny kunnskap om effekten av en behandling. Uten metaanalysen kunne det virke som om MI ikke hadde effekt på rusmiddelbruk – det var jo bare to av tolv studier som viste en signifikant effekt. Men metaanalysen viste at studiene, når de ble slått sammen faktisk viste en signifikant effekt av MI.

REFERANSER

1. Glass GV. Primary, secondary, and meta-analysis of research. *Educational Researcher* 1976; **5**: 3-8.
2. Smedslund G, Berg RC, Hammerstrøm KT, Steiro Ar, Leiknes KA, Karlsen K. Motivational interviewing for substance abuse. *Cochrane Database of Systematic Reviews* 2011; Issue 5. Art.No.:CD008063. DOI: 10.1002/14651858.CD008063.pub2.
3. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1988.