TØI rapport 692/2003
Forfatter: Peter Christensen
Oslo 2003, 78 sider

Sammendrag:

Emner fra meta-analyse
En litteraturstudie

Rapporten oppsummerer en litteraturstudie av metoder for meta-analyse. Formålet med litteraturstudien har vært å dekke felter innen meta-analyse hvor det fremdeles er uløste problemer eller hvor det er uenighet om hvordan meta-analyser bør gjennomføres. Feltene som har pekt seg ut er hvordan man skal behandle heterogene resultater, problemet med publikasjonsskjevhet og hvordan de enkelte undersøkelsers kvalitet skal vurderes og tas hensyn til.

Hva er meta-analyse?

Ofte finner ulike undersøkelser forskjellige virkninger av samme tiltak. En meta-analyse er en beregning av gjennomsnittvirkningen av tiltaket med en systematisk metode for innhenting og bearbeiding av informasjon for sikre at data er så komplette som mulig og for å unngå feilkonklusjoner.

La θ være den sanne virkning av et tiltak. Estimatet for virkningen i den i'te undersøkelsen betegnes med $\hat{\theta}_i$ og variansen av estimatet med $\hat{\sigma}_i^2$. Hvis den sanne variansen σ_i^2 til estimatet er kjent er et optimalt estimat for den sanne virkning θ is gitt ved: $\hat{\theta} = \frac{\sum_{i=1}^{k} w_i \hat{\theta}_i}{\sum_{i=1}^{k} w_i}$, hvor k er antall undersøkelser og $w_i = \frac{1}{\sigma_i^2}$.

Siden σ_i^2 er ukjent brukes estimatet $\hat{\sigma}_i^2$ i vektene, $w_i = \frac{1}{\hat{\sigma}_i^2}$, og usikkerheten i $\hat{\sigma}_i^2$ blir normalt sett bort fra og vektene behandlet som om den sanne σ_i^2 er kjent. Dette er ikke noen alvorlig feil når undersøkelsene som er med i meta-analysen er forholdsvis store.

Fordi parameteren θ is antatt å være den samme (fixed) i alle undersøkelsene kalles dette "fixed effects" metoden.

Heterogenitet

Fixed effects metoden beskrevet ovenfor antar at alle undersøkelser estimerer den samme sanne virkning, dvs at variasjonen i verdier mellom undersøkelser ikke er større enn at den kan forklares ved usikkerheten i de enkelte undersøkelsene. Variasjonen i resultatene kan imidlertid være for stor til at den kan forklares ved standardavviket til resultatene i undersøkelsene. Slik variasjon i resultatene kalles heterogenitet (Thompson and Sharp, 1999), eller mer presist statistisk heterogenitet.

Rapporten kan bestilles fra:
Transportøkonomisk institutt, Postboks 6110 Etterstad, 0602 Oslo
Telefon: 22 57 38 00 Telefax: 22 57 02 90

Om det er heterogenitet eller ikke er ikke alltid åpenbart. Det vil alltid være noe variasjon i resultatene og spørsmålet er om denne er for stor til å forklares med usikkerheten i de enkelte undersøkelsene. Det er derfor utviklet metoder for å påvise og måle heterogenitet.

Heterogenitet kan påvises både ved bruk av grafiske metoder og statistiske tester.

Påvisning og måling av heterogenitet

Grafiske metoder for undersøke om det er heterogenitet er "skogdiagram" (forest plot), Galbraith diagram, en grafisk metode utviklet av Baujat et al og L´Abbé diagram . Disse er beskrevet i hovedteksten.

Standardtesten for heterogenitet er Cochrans Q-test. Test-observatoren er uttrykt ved:

$Q = \sum w_i (\hat{\theta}_i - \hat{\theta})^2$ hvor $\hat{\theta}$ er fixed effect gjennomsnittsestimatet av virkningen som er beskrevet ovenfor og $\hat{\theta}_i$ is den estimerte virkning i undersøkelse i. Q er tilnærmet kjikvadratfordelt.

Ifølge Higgins og Thompson (2002), er det velkjent at testen har liten styrke når det, som vanlig er, er få undersøkelser. Imidlertid er ikke andre tester noe bedre. Takkouche, Cadarso-Suarez, and Spiegelman (1999) undersøkte både type I feil og styrken til Cochrans Q test og fire andre tester for heterogenitet ved simulering. De konkluderte med at med hensyn på validitet, styrke og hvor enkelt det er å beregne testobservatoren, er Cochrans Q den beste. De dårlige nyhetene, som de uttrykker det, er at for de utvalgsstørrelser som er vanlig i epidemiologiske meta- analyser har ingen eksisterende tester akseptable styrke, hvis ikke heterogeniteten er betydelig.

Higgins and Thompson (2002) innfører tre ulike mål for å kvantifisere heterogenitet. Et av de to målene som de anbefaler er basert på Cohrans Q og er gitt ved:

$$H^2 = \frac{Q}{k-1}$$

Meta-analyse når det er heterogenitet

Fixed effects metoden tar ikke hensyn til heterogenitet. Når det er heterogenitet vil derfor fixed effects metoden underestimere usikkerheten i det veiede gjennomsnittestimatet. En metode som tar hensyn til den økte usikkerheten er derfor nødvendig.

Dessuten vil ikke en beregning av det veide gjennomsnittestimatet når det er heterogenitet forklare heterogeniteten. Er den reell eller har den metodologiske forklaringer? Under hvilke forhold virker tiltaket? En forklaring på heterogeniteten er også av interesse. Det er derfor to ulike analyttiske angrepsmåter når det gjelder heterogenitet, å bare ta hensyn til den eller å forsøke å forklare den.

Hvordan ta hensyn til heterogenitet

Random effects (RE) metoden tar hensyn til heterogenitet men forklarer ikke hvorfor den oppstår. RE metoden bygger på en antagelse om at den sanne virkning θ_i i den i'te undersøkelsen er tilfeldig valgt fra en normalfordeling av undersøkelser med gjennomsnittlig virkning θ. Mer presist bygger RE metoden på følgende modell. Den observerte virkning x_i i den i'te undersøkelsen er gitt ved:

$x_i = \theta_i + e_i$ hvor $E(e_i)=0$ og $Var(e_i)=\sigma_i^2$ og $\theta_i = \theta + u$, $E(u)=0$ og $Var(u)=\tau^2$.

σ_i^2 er variansen til resultatet i en undersøkelse og τ^2 er variansen mellom undersøkelser. $Var(x_i)$ er nå gitt ved $Var(x_i)= \sigma_i^2 + \tau^2$ og vektene i fixed effects metoden er erstattet

med random effects vekter $w_i^* = \frac{1}{\sigma_i^2 + \tau^2}$. Bruk av disse vektene krever estimater for σ_i^2 og τ^2.

Som nevnt tidligere for fixed effects metoden, så er σ_i^2 antatt å være kjent, dvs at estimatene $\hat{\sigma}_i^2$ for σ_i^2 er antatt å være uten usikkerhet. Denne antagelsen er rimelig hvis antall observasjoner eller tilfeller i undersøkelsene er stort. En tilsvarende antagelse for estimatet av τ^2 er mindre rimelig siden antall undersøkelser i en meta-analyse vanligvis er ganske lite. Vanligvis vil usikkerheten i estimatet for τ^2 være betydelig. Likevel tar ikke det mest vanlige estimatet for τ^2 og den mest vanlige varianten av random effects analyse hensyn til usikkerheten til estimatet til τ^2. I dette tilfellet er variansen til det veide gjennomsnittet gitt ved:

$$\frac{1}{\sum \frac{1}{\hat{\sigma}_i^2 + \tau^2}}.$$

Det langt vanligste estimatet for τ^2 er DerSimonian og Laird estimatet som bygger på Cochran's Q-test. Det er en moment-basert estimator som fås ved å sette den observerte verdien av Q lik forventningen og er gitt ved:

$$\tau_{DL}^2 = \frac{Q - (k-1)}{\sum w_i - \frac{\sum w_i^2}{\sum w_i}}.$$

Når Q<k-1, settes τ_{DL}^2 lik null. τ_{DL}^2 er derfor et oppad avrundet estimat og følgelig forventningsskjevt.

Denne verdien av τ^2 brukes i vektingsformelen over og det veide gjennomsnittet kan beregnes.

Alternativt kan beregnes en sannsynlighetsmaksimeringsestimator. I dette tilfellet beregnes τ^2 og det veide gjennomsnittet $\hat{\theta}$ simultant. Ligningene for løsning er gitt i Hardy og Thompson (1996).

$$\hat{\theta} = \frac{\sum \frac{\hat{\theta}_i}{(\hat{\sigma}_i^2 + \hat{\tau}^2)}}{\sum \frac{1}{(\hat{\sigma}_i^2 + \hat{\tau}^2)}} \quad \text{og} \quad \hat{\tau}^2 = \frac{\sum \frac{(\hat{\theta}_i - \theta)^2 - \sigma_i^2}{(\hat{\sigma}_i^2 + \hat{\tau}^2)^2}}{\sum \frac{1}{(\hat{\sigma}_i^2 + \hat{\tau}^2)^2}}.$$ Det ses at ligningen for $\hat{\theta}$ er standard random effects Verdi for $\hat{\theta}$ gitt $\hat{\tau}^2$.

Ligningene løses ved iterasjon. Med en startverdi for τ^2 kan finnes en løsning for $\hat{\theta}$. Brukes den verdien i den andre ligningen kan fås en ny verdi for τ^2, osv. Hvis dette estimatet brukes uten å ta hensyn til usikkerheten i estimatet, vil det bli betegnet som enkel sannsynlighetsmaksimering.

Hardy og Thompson (1996) bruker profil sannsynlighetsmaksimering til å konstruere konfidensintervaller. Den følgende beskrivelsen er hentet fra deres artikkel.

Profil log-likelihood i to-parameter tilfellet er log-likelihood for én parameter gitt et estimat for den andre, dvs $l_1^*(\theta) = l(\theta, \hat{\tau}^2(\theta))$ og $l_2^*(\tau^2) = l(\hat{\theta}(\hat{\tau}^2), \tau^2))$. $\hat{\tau}^2(\theta)$ is sannsynlighetsmaksimeringsestimatoren for τ^2 som en funksjon av

θ og $\hat{\theta}(\tau^2)$ er sannsynlighetsmaksimeringsestimatoren of θ som en funksjon av τ^2. Et konfidensintervall for τ^2 er gitt ved verdiene som tilfredstiller $l_2^*(\tau^2) > l_2^*(\hat{\tau}^2) - 3.84/2$. Konfidensintervallet er ikke nødvendigvis symmetrisk.

Å forklare heterogenitet

Flere forfattere understreker betydningen av å forklare heterogenitet og ikke bare ta hensyn til den ved å bruke random effect-metoden. Den anbefalte måten å gjøre dette på er å bruke meta-regresjon. Betegnelsen meta-regresjon brukes når de uavhengige variablene beskriver egenskaper ved undersøkelsene, i motsetning til de regresjonsanalyser som er mulige når data på individnivå er tilgjengelige fra hver undersøkelse (Thompson and Higgins, 2002).

Meta-regresjon har to viktige egenskaper. For det første, siden undersøkelsene som er enhetene i analysen sjelden vil være av samme størrelse, og variansene til estimatene derfor er forskjellige, er det heteroskedastisitet, og vektet regresjon er nødvendig. For det andre, det er lite sannsynlig at regresjon vil forklare all heterogenitet og den gjenværende heterogenitet må tas hensyn til i den statistiske analysen. Den korrekte regresjonsmodell er derfor en random effects modell hvor vekten for hver undersøkelse er lik den inverse av summen av undersøkelsenes varians og den gjenværende varians mellom undersøkelser, analogt med random effects modellen beskrevet ovenfor.

Den gjenværende variansen mellom undersøkelser er bare kjent etter at regresjonsanalysen er gjennomført. En metode for å estimere regresjonsligningen og τ^2 samtidig eller med iterasjon er derfor nødvendig. Thompson and Sharp (1999) beskriver fire metoder. De er beskrevet i hovedteksten.

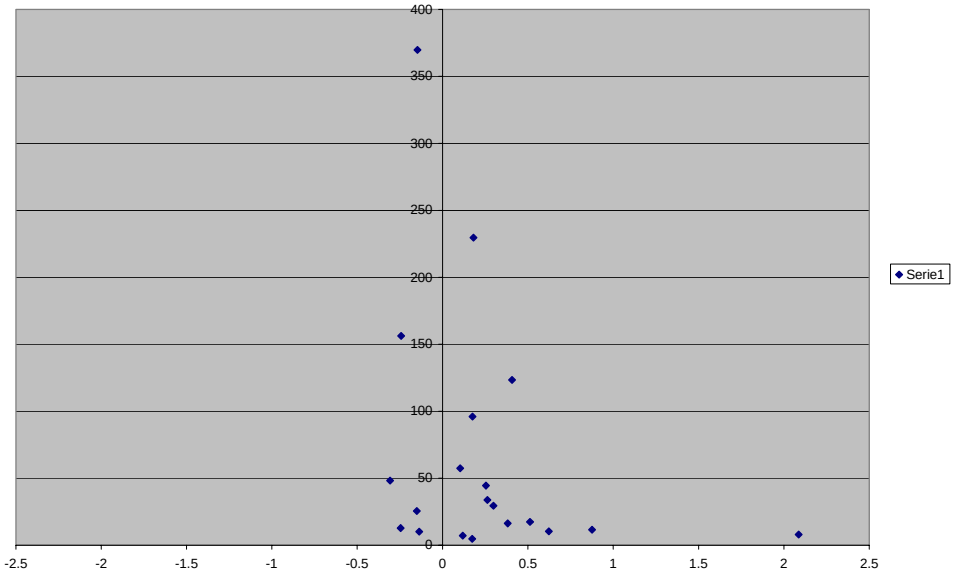
Publikasjonskjevhet

Hvis undersøkelser som publiseres skiller seg fra undersøkelser som ikke er publisert med hensyn på hvilken virkning som er funnet, dvs at resultatene påvirker sannsynligheten for publisering, vil publiserte undersøkelser være et skjevt utvalg av alle undersøkelser. Dette er *publikasjonsskjevhet*.

Flere undersøkelser har vist at publikasjonsskjevhet er vanlig. Noen av disse er beskrevet i hovedteksten. Hovedtemaet i denne rapporten er imidlertid metoder for å undersøke om undersøkelsene som inngår i en meta-analyse er påvirket av publikasjonsskjevhet.

Et verktøy for å undersøke mulig publikasjonsskjevhet er traktdiagrammet. I et traktdiagram plottes virkningene som er funnet i et utvalg undersøkelser mot et mål for resultatenes presisjon.

Et eksempel på et traktdiagram er vist i figur 1.



Figur 1. Traktdiagram. Eksempel basert på vilkårlige data. TØI rapport 692/2003.

I figuren er virkningen uttrykt ved logaritmen til oddsforholdet og presisjonen er undersøkelsenes vekt, dvs den inverse til variansen til logaritmen til oddsforholdet.

Årsaken til betegnelsen traktdiagram er at når det ikke er publikasjonsskjevhet vil diagrammet se ut som en trakt snudd på hodet. Store undersøkelser vil være mer nøyaktige og vise mindre variasjon enn små undersøkelser. Når det ikke er publikasjonsskjevhet vil diagrammet være symmetrisk om gjennomsnittet.

Et forhold som skaper publikasjonsskjevhet er hvis resultater som ikke er signifikante ikke blir publisert. Siden små undersøkelser krever en større observert virkning for å være statistisk signifikant vil det være en tendens til at man finner større virkninger for små undersøkelser fordi små undersøkelser som finner liten eller endog en negativ virkning vil mangle. Det vil da mangle undersøkelser nede til venstre i traktdiagrammet. Dette ser ut til å være tilfellet i traktdiagrammet ovenfor og kan tolkes som et tegn på publikasjonsskjevhet.

Traktdiagrammet utnytter forskjellen i virkning mellom store og små undersøkelser. For at et traktdiagram skal være egnet er det derfor nødvendig at undersøkelsene varierer i størrelse.

For testing av publikasjonsskjevhet finnes statistiske metoder som er analoge til traktdiagrammet. Tre metoder beskrives her. To av metodene bygger på antagelsen om at publikasjonsskjevhet leder til en sammenheng mellom størrelsen på virkningen som er funnet og standardavviket til virkningen. En av metodene tester for en slik sammenheng ved bruk av en rangkorrelasjon test og den andre bruker regresjonsanalyse. Den tredje metoden tester for symmetri i traktdiagrammet.

Beggs test (Begg, 1994) er en test for uavhengigheten mellom størrelsen og variansen av virkningen og bygger på Kendalls tau. Testen er basert på en antagelse om at størrelsen av virkningene er statistisk uavhengige og identisk fordelte under nullhypotesen om at det ikke er publikasjonsskjevhet. Det er derfor nødvendig å standardisere størrelsen av virkningene for å kunne utføre testen. Betegnes størrelsen av virkningen og dens varians

med henholdsvis x_i og v_i, brukes rangkorrelasjon til å teste sammenhengen mellom $x_i = \frac{(x_i - \bar{x})}{\bar{v}_i^{\frac{1}{2}}}$,

hvor $\bar{x}$ er fixed effects gjennomsnitt av virkningene og hvor $\bar{v} = v_i - \left(\sum_{j=1}^{n} v_i^{-1} \right)^{-1}$ er variansen til $x_i - \bar{x}$.

Testen går ut på å beregne P, antallet av alle mulige par hvor de tofaktorene er rangert i samme rekkefølge , og Q, antall par hvor de ikke er i rekkefølge. En normalisert test observator (z score) er da gitt ved:

$$Z = \frac{(P - Q)}{\left[n(n-1)(2n+5)/18 \right]^{\frac{1}{2}}}$$

Egger m fl (1997) bruker lineær regresjon til undersøke sammenhengen mellom virkningens størrelse og varians og på den måten teste for publikasjonsskjevhet. I regresjonen er den avhengige variabelen den standardiserte virkningen og den uavhengige variabelen er den inverse av standardavviket til virkningen. Betegnes virkningen med x og standardavviket med s er regresjonsligningen:

$$\frac{x}{s} = a + b\frac{1}{s}$$

Testen for publikasjonsskjevhet er basert på koeffisienten *a*. En signifikant verdi tyder på publikasjonsskjevhet.

Begrunnelsen for testen er følgende. Den inverse av standardavviket er et mål for presisjon. Unøyaktige undersøkelser (vanligvis små undersøkelser) vil ligge nær origo på abscissen. Den standardiserte virkning vil da også være liten. Unøyaktige undersøkelser vil derfor ha små verdier på begge akser, dvs de vil ligge nær origo. Nøyaktige undersøkelser vil ligge langt fra origo på abscissen og hvis der er en virkning vil ordinatverdien også være stor. Regresjonslinjen gjennom de plottede undersøkelsene vil derfor gå tilnærmet gjennom origo med en vinkelkoeffisient lik den veide gjennomsnittsvirkningen. Dette er situasjonen når det ikke er publikasjonsskjevhet.

Når traktdiagrammet er asymmetrisk fordi det er publikasjonsskjevhet og mindre undersøkelser finner virkninger om avviker systematisk fra store undersøkelser vil ikke regresjonsligningen gå gjennom origo. Konstantleddet, koeffisienten *a*, vil derfor være et mål for asymmetrien. Fortegnet til *a* vil avhenge av hvordan virkningen måles. Når virkningen måles ved logaritmen til oddsforholdet og en negative verdi betyr en positiv virkning vil koeffisienten være negativ når det er publikasjonsskjevhet. En test for publikasjonsskjevhet fås følgelig ved å teste om koeffisienten *a* er forskjellig fra null. Da testen har lav styrke anbefaler Egger m fl å bruke er 10 % signifikansnivå.

"Trim and fill"-metoden til Duval and Tweedie (2000a, 2000b) bygger også på traktdiagrammet men i dette tilfellet er ikke utgangspunktet sammenhangen mellom virkningens størrelse og dens varians men symmetrien (eller mangel på symmetri) i traktdiagrammet.

Hvis det ikke er publikasjonsskjevhet, vil traktdiagrammet være symmetrisk. "Trim and fill"-metoden fjerner derfor et antall undersøkelser på en side av diagrammet slik at det blir symmetrisk (trimdelen), beregner et veid gjennomsnitt av de resterende undersøkelsene og genererer så et tilsvarende antall undersøkelser på den andre siden. De genererte undersøkelsene er symmetrisk til de fjernede undersøkelsene rundt det beregnete gjennomsnittet. Et eksempel med diagrammer er vist i hovedteksten.

Kvalitetsvurdering og kvalitetscore

Bangert-Drowns, Wells-Parker and Chevillard (1997) påpeker at hvis kvaliteten på undersøkelsene ikke tas hensyn til, så betyr dette at undersøkelser med store utvalg er overlegne andre undersøkelser på grunnlag av bare denne egenskapen; utvalgsstørrelse.

Den eneste usikkerhet ved undersøkelsene som det tas hensyn til er den statistiske, som om det ikke er noen metodiske problemer. Dette er selvfølgelig meget urealistisk og er alene et argument for å vurdere kvaliteten på undersøkelsene.

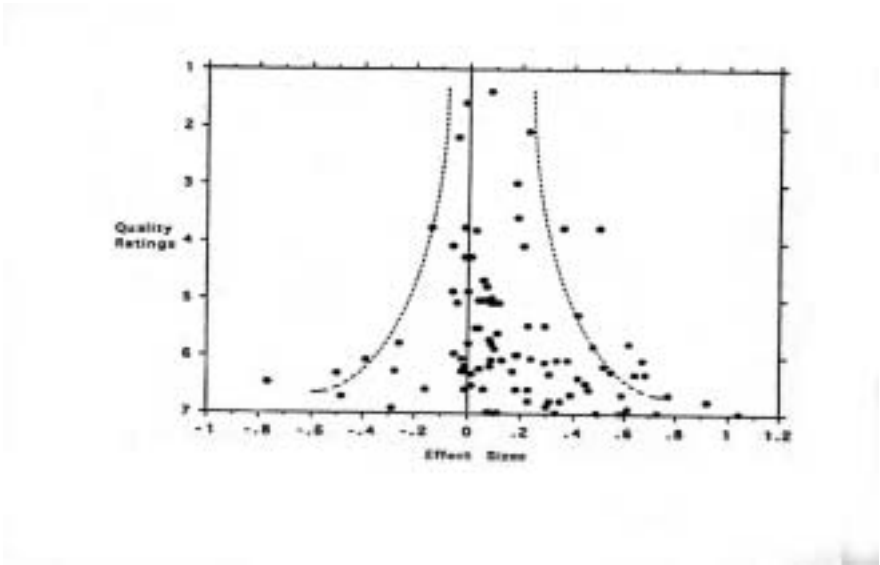
Bruk av kvalitetsvurderinger i meta-analyser krever svar på to spørsmål, hvordan kvalitet skal måles og hvordan kvaliteten på undersøkelsene skal tas hensyn til i meta-analysen. Svaret på siste spørsmål avhenger av de mulige virkninger av dårlig kvalitet av undersøkelser. Virkningen kan være:

1. En systematisk skjevhet
2. Økt varians (en større usikkerhet eller mindre presisjon)

Flere undersøkelser har funnet at undersøkelser av lav kvalitet har en tendens til å overvurdere virkningen. Imidlertid har andre undersøkelser ikke funnet at undersøkelser med lav kvalitet leder til systematiske skjevheter. Balk et al (2002) gjennomførte en empirisk undersøkelse av korrelasjonen mellom ulike kvalitetsmål og estimatet av virkningen av behandling. 24 kvalitetsmål ble analysert for 276 randomiserte kontrollerte forsøk fra 26 meta-analyser. Kvalitetsmålene ble omkodet til dikotome variable, høy kvalitet eller lav kvalitet. Betydningen av kvalitetsmålet ble estimert ved å beregne det relative oddsforholdet for hvert mål. Relative oddsforhold for undersøkelser med høy kvalitet mot undersøkelser med lav kvalitet varierte for de ulike kvalitetsmålene mellom 0.83 og 1.26. Ingen hadde signifikant sammenheng med størrelsen på virkningen.

Hvis resultatene fra undersøkelser av lav kvalitet varierer mer enn resultatene fra gode undersøkelser, fordi resultatene fra dårlige undersøkelser er mindre pålitelige, vil kvalitet ha samme virkning på variasjonen i resultatene som utvalgsstørrelse. Et plot av størrelsen på virkningen mot kvalitet vil oppføre seg på samme måte som et traktdiagram med størrelsen på virkningen mot statistisk vekt.

Dette er vist i figur 2 hentet fra Bangert-Drowns, Wells-Parker og Chevillard (1997). Lave tall betyr høy kvalitet.



Figur 2. Punktdiagram for sammenhengen mellom virkningen på residivisme og vurderinger av metodologiisk kvalitet av undersøkelser av rehabiliteringsprogrammer for promilleførere.
Fra Bangert-Drowns, Wells-Parker og Chevillard (1997).

Hvordan ta hensyn til kvalitet

Den større usikkerheten til dårlige undersøkelser underbygger behovet for å ta hensyn til undersøkelsenes kvalitet på en eller annen måte. Det finnes (minst) fem måte å gjøre dette:

1. Utelate de dårligste undersøkelsene, dvs definere en terskel for kvalitet og bare ta med undersøkelser over denne terskelen i meta-analysen.
2. Stratifisere undersøkelsene med hensyn på kvalitet og gjøre en separat meta-analyse for hvert stratum.
3. Bruke kvalitetscore som vekter på same måte som man nå bruker statistiske vekter.
4. Bruke meta-regresjon til å uttrykke virkningen som en funksjon av enten kvalitet eller komponentene som inngår i kvalitetsmålet
5. Sekvensiell sammenveiing resultatene fra undersøkelsene basert på kvalitetscore (Detsky et al, 1992).

Ingen av de to første metodene er tilfredstillende. Enhver terskel vil nødvendigvis være vilkårlig. I tillegg vil ikke kvaliteten lenger være av betydning for de undersøkelser som ligger over terskelen. En undersøkelse vil enten ha ingen vekt eller full vekt avhengig av terskelen.

Å stratifisere etter kvalitet løser heller ikke problemet, det er bare utsettelse. Det gir ikke noe svar på spørsmålet om hva som skal gjøres med resultatene fra meta-analysene for hver stratum. Hvis man bare stoler på resultatet fra stratumet med høyest kvalitet er dette ekvivalent med bruk av en terskel. Hvis resultatene fra alle strata skal benyttes gjenstår fremdeles spørsmålet om hvordan resultatene fra de enkelte strata skal veies sammen.

Meta-regresjon bygger på den antagelse at det er en systematisk sammenheng mellom en kvalitetsskala eller komponentene til en kvalitetsskala og resultatene i en undersøkelse. Hvis det ikke er noen systematisk sammenheng vil meta-regresjonen tyde på at de metodologiske variable ikke påvirker forskjellen i resultater mellom undersøkelser. Likevel vil variasjoner som skyldes metodologiske mangler bidra til at størrelsen på feilleddet øker og følgelig blir konfidensintervallene videre. Dårlige undersøkelser vil imidlertid påvirke resultatet like meget som gode undersøkelser.

Sekvensiell sammenveiing av resultatene fra undersøkelsene basert på kvalitetscore kan betraktes som en spesiell form for bruk av en terskel for kvalitet og lider derfor av samme svakhet som bruk av en terskel.

Kvalitetsscoring har få tilhengere. Jüni, Altman and Egger (2001a) mener at "bruk av kvalitetsscore som vekter mangler statistisk og empirisk begrunnelse" og at det er ingen grunn til at kvalitetsscore skal modifisere estimatets presisjon. Problemene med kvalitestscore kan illustreres ytterligere av en undersøkelse av Jüni et al (1999).

De evaluerte bruken av 25 forskjellige vurderingsskalaer beskrevet av Moher m fl (1995). Disse skalaene ble brukt på 17 forsøk som sammenlignet bruk av ulike hepariner for å forebygge blodpropp etter operasjoner.

Selv om overenstemmelsen for standardiserte scorer for de 25 skalaene var betydelig (intraclass korrelasjonkoeffisient 0.72 (0.59,0.86)) varierte medianen av av kvaliteten til undersøkelsene fra 38.5% til 82.9% av maksimum score for de ulike skalaene. Ved bruk av kvalitetscore som vekter i en meta-analysis, ville konfidensintervaller basert på ska-

laen med den laveste median score være mer enn to ganger så vide som konfidensintervall basert på skalaen med den høyeste median score.

Resultatet ser ut til å være et argument mot å bruke kvalitetsscore i meta-analyser. Vårt syn er imidlertid at resultatet bare gjenspeiler vilkårligheten i eksisterende kvalitetsscore og kan unngås ved å utvikle bedre kvalitetsskalaer.

Måling av kvalitet

Ulike kvalitetsvurderinger vil avvike dersom de underliggende kvalitetsbegrepene er forskjellige. Arbeidet med å vurdere kvalitet må derfor begynne med å avgrense hva som skal ligge i begrepet "kvaliteten av en undersøkelse".

I hvilken sammenheng begrepet skal brukes er viktig for definisjonen av kvalitet. En vurdering av kvaliteten av en undersøkelse innsendt til et tidskrift for publisering er noe annet en vurderingen av kvaliteten til an undersøkelse når den inngår i en meta-analyse. I det første tilfellet er kvalitetsbegrepet meget videre enn i det siste, det omfatter f eks interesse for leserne, resultatenes orginalitet osv. En undersøkelse som inngår i en meta-analyse kan betraktes som et instrument for å måle virkningen av noe. Kvaliteten til undersøkelsen er et mål for i hvilken grad man kan stole på resultatene, dvs undersøkelsens reliabilitet og validitet.

Dette leder til følgende definisjon av kvalitet: Den grad en undersøkelse har unngått metodologiske svakheter som kan påvirke resultatene. Dette er nesten (men ikke helt) det samme som begrepet "intern validitet" hos Shadish, Cook and Campbell (2002). Noen forfattere (Downs and Black 1998, Verhagen et al 1998) er av den oppfatning at kvalitetsbegrepet også bør omfatte ekstern validitet. Vårt syn er at ekstern validitet må tas hensyn til gjennom meta-analysen ved å sikre at de inkluderte undersøkelsene varierer med hensyn på enheter som inngår i undersøkelsene og forholdene de blir utført under. Meta-regresjon kan så benyttes til å analysere virkningen av variasjonen. Ekstern validitet bør ikke inngå i definisjonen av kvalitet.

Jadad et al (1996) nevner tre metoder for å vurdere kvaliteten til kliniske forsøk: individuelle markører, også kalt elementer eller komponenter, sjekklister og skalaer.

Individuelle markører er de mulige dimensjonene til kvalitet. Eksempler for randomiserte kontrollerte forsøk er randomiseringsmetoden og blindingsmetoden. For kvasi-eksperimentelle undersøkelser er det vanskeligere å gi eksempler på individuelle markører.

Sjekklister gir et kvalitativt anslag for kvaliteten til en undersøkelse med utgangspunkt i individuelle markører eller komponenter for å sammenligne undersøkelser (Moher, Jadad and Tugwell, 1996). For sjekklister beregnes ikke noe numerisk score.

En skala konstrueres ved å gi komponentene numeriske verdier og summere (hvis ønskelig veid) disse. For å vekte undersøkelser er kvalitetsskalaer å foretrekke.

Med den gitte definisjonen av kvalitet må kvalitetsskalaen være et mål for hvor godt faktorer som kan påvirke resultatet har blitt kontrollert for. En måte å kode undersøkelser er derfor å rangere ulike design for deres evne å kontrollere for bakgrunnsvariable og betrakte kvaliteten på undersøkelsen høyere jo lavere undersøkelsdesignet er rangert. Hvor gjennomført designet er implementert bør også tas hensyn til. Et alternativ er å lage en oversikt over mulige bakgrunnsvariable og å sjekke hvor godt undersøkelsen har kontrollert for dem. Jo flere bakgrunnsvariable som er kontrollert for jo høyere er kvaliteten på undersøkelsen.

Å evaluere validiteten til en kvalitetsskala er vanskelig. For å verifisere den teoretiske validiteten til en kvalitetsskala er det nødvendig å avlede konsekvenser av kvalitet som kan bekreftes empirisk. En slik konsekvens diskutert ovenfor er at virkningene funnet i undersøkelser av god kvalitet vil variere mindre enn virkningene funnet i undersøkelser

av dårlig kvalitet. Dette kan testes i et traktdiagram med virkningen langs abscissen og kvalitetsscoren langs ordinaten.

Det er to vesentlige forutsetninger for bruken av kvalitetsvurderinger i meta-analyser:

1. En undersøkelses kvalitet bør ha betydning for dens vekt i meta-analysen
2. Usikkerheten som skyldes metodologiske svakheter bør gjenspeiles i det veide estimatet for virkningen.

For å oppnå dette er en kvalitetskala nødvendig.