



Kan vi stole på subgruppeanalyser?

MEDISIN OG TALL

EVA SKOVLUND

E-post: eva.skovlund@ntnu.no

Eva Skovlund er professor i medisinsk statistikk ved Institutt for samfunnsmedisin og sykepleie, NTNU, og seniorforsker ved Folkehelseinstituttet.

Forfatteren har fylt ut ICMJE-skjemaet og oppgir ingen interessekonflikter.

Funn i subgruppeanalyser fra kliniske forsøk er ofte omdiskuterte. Hvordan kan vi best skille mellom tilfeldige funn og faktiske forskjeller i effekt når det ser ut til at en gruppe pasienter har bedre effekt av behandlingen enn en annen?

I kliniske studier ønsker man ofte ikke bare å studere om en behandling er bedre enn en annen, men også å besvare følgende spørsmål: Hvilke pasienter vil ha nytte av behandlingen? Analyser av subgrupper av pasienter kan kanskje gi et svar, og resultatet kan ha stor betydning for klinisk praksis. Men slike analyser kan medføre høy risiko for falskt positive funn, og det finnes mange eksempler på at man har trukket feilaktige konklusjoner (1, 2).

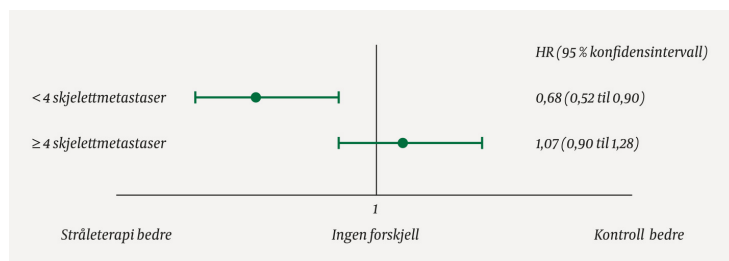
Interaksjonstest

Det er ikke uvanlig at det i kliniske studier rapporteres en rekke subgruppeanalyser og en separat p-verdi for hver analyse (1). Anta at man vil undersøke om pasientens allmenntilstand har betydning for effekt av behandling, og at man finner en statistisk signifikant effekt ($p < 0,001$) i subgruppen med dårlig allmenntilstand, men ikke hos pasienter med god allmenntilstand ($p = 0,22$). Enkelte lar seg friste til å sammenligne p-verdiene og konkludere med at det kun er effekt i gruppen med dårlig allmenntilstand. Dette er ikke en anbefalt strategi, blant annet fordi p-verdier avhenger sterkt av antallet pasienter som er inkludert i analysen. En p-verdi som er høyere enn 0,05, betyr ikke nødvendigvis at det ikke er forskjell i effekt, og hvis den ene subgruppen er stor og den andre liten, vil p-verdien ofte bli «signifikant» kun i den første. Det viktige er ikke p-verdiene, men estimatet av effekt i hver gruppe samt den tilhørende usikkerheten.

Den korrekte tilnærmingen til å studere om effekt av behandling avhenger av gitte pasientkarakteristika, er en interaksjonstest (3, 4). Med en slik test sammenligner man behandlingseffekt i de ulike subgruppene. En lav p-verdi betyr at behandlingseffekten i en subgruppe er signifikant forskjellig fra effekten i den (eller de) andre subgruppen(e). Da gir det også mening å presentere separate effektestimater, inkludert 95 % konfidensintervall, i hver av subgruppene. Dette gir vesentlig mer informasjon enn kun en p-verdi. Merk at resultatet er skalaavhengig, og at vi for eksempel kan ha signifikant interaksjon på relativ skala, men ikke på absolutt skala (4).

Et eksempel

I en randomisert studie ved prostatakreft fant man ingen forskjell i total overlevelse mellom gruppen som fikk stråleterapi, og kontrollgruppen (HR = 0,92, 95 % konfidensintervall 0,80 til 1,06) (5). Men i en forhåndsspesifisert subgruppeanalyse der pasientene ble delt inn i to grupper – de med tre eller færre skjelettmetastaser og de med fire eller flere – fant man forskjell i gruppen med få skjelettmetastaser ($p(\text{interaksjon}) = 0,0098$) (figur 1). Denne analysen var en av to subgruppeanalyser som var forhåndsspesifisert i protokollen, og hypotesen om heterogenitet var basert på publiserte funn i en tidligere studie. Dette gjør funnet mer troverdig enn om det hadde vært basert på en post hoc-analyse.



Figur 1 Effekt av behandling i to subgrupper (5). Det er statistisk signifikant bedre effekt av stråleterapi i subgruppen med færre enn fire skjelettmetastaser ($p = 0,0098$).

Legg merke til at ingen av de to konfidensintervallene i figur 1 overlapper punkttestimatet av hasardratioen (HR) i den andre subgruppen. Det gir en indikasjon på at forskjellen mellom subgruppene neppe skyldes tilfeldighet. Det er dette som viser at effekten er forskjellig, ikke at det ene konfidensintervallet inneholder verdien 1, mens det andre indikerer redusert risiko for død.

Troverdighet av funn

Et generelt problem med subgruppeanalyser er at risikoen for falskt positive funn øker med antallet signifikanstester som utføres. Dette er problematisk hvis man ønsker å studere mange subgrupper. Et annet problem er at subgrupper ofte er små, og at teststyrken dermed blir lav. Man bør begrense antallet subgruppeanalyser slik at man unngår høy sannsynlighet for falskt positive funn, og analysene bør være forhåndsspesifisert dersom man skal benytte dem til å trekke en konklusjon. En post hoc-analyse er mer egnet til å generere en hypotese enn til å trekke en slutning om effekt. I tillegg er det viktig å vurdere biologisk «rimelighet» og sjekke om et resultat følger samme mønster for beslektede utfall (2).

Når subgruppeanalyser er planlagt, utført og rapportert på en edruelig måte, kan de gi verdifull informasjon om hvilke pasientgrupper som har nytte av behandling. Men antallet subgruppeanalyser i en studie bør være lite og spesifisert i studieprotokollen, og en eventuell konklusjon om heterogenitet skal baseres på en interaksjonstest, ikke på sammenligning av separate p-verdier.

LITTERATUR:

1. Hernández AV, Boersma E, Murray GD et al. Subgroup analyses in therapeutic cardiovascular clinical trials: are most of them misleading? *Am Heart J* 2006; 151: 257–64. [PubMed][CrossRef]
2. Sun X, Ioannidis JP, Agoritsas T et al. How to use a subgroup analysis: users' guide to the medical literature. *JAMA* 2014; 311: 405–11. [PubMed][CrossRef]
3. Wang R, Lagakos SW, Ware JH et al. Statistics in medicine—reporting of subgroup analyses in clinical trials. *N Engl J Med* 2007; 357: 2189–94. [PubMed][CrossRef]
4. Thoresen M. Er effekten forskjellig blant kvinner og menn? *Tidsskr Nor Lægeforen* 2018; 138. doi: 10.4045/tidsskr.18.0602. [PubMed][CrossRef]

5. Parker CC, James ND, Brawley CD et al. Radiotherapy to the primary tumour for newly diagnosed, metastatic prostate cancer (STAMPEDE): a randomised controlled phase 3 trial. *Lancet* 2018; 392: 2353–66. [PubMed][CrossRef]

Publisert: 4. mai 2020. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.20.0119
© Tidsskrift for Den norske legeforening 2020. Lastet ned fra tidsskriftet.no