TØI-rapport 984/2008
Forfatter(e): Rune Elvik
Oslo 2008, 140 sider

Sammendrag:

Vurdering av kvaliteten på undersøkelser om virkninger av trafikksikkerhetstiltak

Denne rapporten presenterer en undersøkelse av muligheten for å utvikle et systematisk opplegg for å vurdere kvaliteten på undersøkelser om virkninger av trafikksikkerhetstiltak. Rapporten utgjør den siste rapporten fra det strategiske instituttprogrammet "Bruk av meta-analyser til kunnskapsoppsummering i transportforskning", som formelt pågikk fra 2000 til 2004.

Bakgrunn og problemstilling

Omfanget av forskning øker på nesten alle fagområder og det er en stor utfordring å sammenfatte foreliggende kunnskap på en konsis og riktig måte. Ett av problemene man møter på mange fagområder, er at kvaliteten på foreliggende undersøkelser varierer. Man ønsker da å legge mest vekt på de beste undersøkelsene. Dette krever at man kan bedømme kvaliteten på undersøkelser på en systematisk måte.

Det foreligger i dag flere tusen studier om virkninger av trafikksikkerhetstiltak. Mange av disse studiene er oppsummert i Trafikksikkerhetshåndboken, som er under kontinuerlig oppdatering og utvikling. I Trafikksikkerhetshåndboken presenteres til dels svært detaljerte opplysninger om virkninger av mange trafikksikkerhetstiltak, noe som kan gi inntrykk av at det foreligger omfattende kunnskap om virkninger av slike tiltak.

Det er riktig at det er utført omfattende forskning om virkninger av trafikksikkerhetstiltak, men dessverre er ikke all denne forskningen av like god kvalitet. Det er lett å finne eksempler på dårlige undersøkelser, og det er lett å vise eksempler på hvordan svakheter ved de dårlige undersøkelsene har påvirket resultatene av dem. I rapporten gis en rekke slike eksempler.

På denne bakgrunn, er hovedproblemstillingen denne rapporten tar sikte på å besvare:

Er det mulig å bedømme kvaliteten på undersøkelser om virkninger av trafikksikkerhetstiltak på en systematisk måte, fortrinnsvis i form av en tallmessig skala for undersøkelsers kvalitet?

Det er tidligere, spesielt i medisin, gjort mange forsøk på å utvikle tallmessige mål på undersøkelsers kvalitet. En tungtveiende innvending mot de aller fleste av disse målene, er at de i stor grad er vilkårlige, det vil si at det i liten grad gis noen begrunnelse av hva som inngår i dem og hvordan ulike poster er vektet i forhold

Rapporten kan bestilles fra:
Transportøkonomisk institutt, Gaustadalléen 21, NO 0349 Oslo
Telefon: 22 57 38 00 Telefax: 22 60 92 00

til hverandre. Kvalitet på undersøkelser er følgelig et begrep det er vanskelig å operasjonalisere på en velbegrunnet måte.

Kan en ikke-vilkårlig skala for kvalitet utvikles?

I denne rapporten er det lagt vekt på å etablere et grunnlag for å utvikle en skala for kvalitet der de ulike elementene i størst mulig grad begrunnes, slik at vilkårligheten reduseres. For å oppnå dette, er flere tilnærmingsmåter valgt:

- Gjennomgang av tidligere utviklede skalaer for tallfesting av undersøkelsers kvalitet,
- Spørreundersøkelse blant ledende trafikksikkerhetsforskere om hva de legger i begrepet kvalitet og om de mener det er mulig å lage et tallmessig mål på undersøkelsers kvalitet,
- Utvikling og testing av en pilotversjon av en skala for kvalitet på undersøkelser om trafikksikkerhetstiltak,
- Utvikling av en typologi av undersøkelsesopplegg i studier av trafikksikkerhetstiltak og mulige feilkilder i slike undersøkelser,
- Gjennomgang av metodologisk forskning om hvilken betydning ulike feilkilder kan ha for resultatene av undersøkelser om virkninger av trafikksikkerhetstiltak.

Tidligere kvalitetsskalaer

35 ulike skalaer som er utviklet for å måle kvaliteten på undersøkelser er gjennomgått. De aller fleste av disse skalaene er utviklet i medisin. Kun et fåtall skalaer for trafikksikkerhetsstudier ble funnet.

De færreste skalaer for kvalitet på undersøkelser bygger på en klar definisjon av begrepet kvalitet. Det varierer svært mye hva som inngår i skalaene, og hele 158 variabler ble kodet for å definere innholdet i de 35 skalaene. Disse 158 variablene kan reduseres til 12 hovedkategorier. Det er imidlertid høyst tvilsomt om alle disse kategoriene har særlig mye med undersøkelsers kvalitet å gjøre. Reliabiliteten er testet for noen av skalaene og har vist seg å være god. Validiteten av skalaene er i liten grad testet, og de få tester som foreligger er til dels meningsløse.

I det hele tatt må det konkluderes med at gjennomgangen av tidligere utviklede skalaer for undersøkelsers kvalitet langt på veg bekrefter den kritikk som har vært reist mot slike skalaer, nemlig at de er vilkårlige, ubegrunnede, subjektive og ikke testet på en vitenskapelig holdbar måte. Praktisk talt ingen ting av nytte for å utvikle en skala for kvalitet på undersøkelser om virkninger av trafikksikkerhets-tiltak kom ut av gjennomgangen av foreliggende skalaer.

Ekspertoppfatninger om kvalitet

Fire åpne spørsmål om kvalitet på undersøkelser om virkninger av trafikksikker-hetstiltak ble sendt til 10 av verdens ledende trafikksikkerhetsforskere. Det kom 8

svar. Svarene viste at oppfatningene om hva som ligger i begrepet kvalitet varierer mye. Det var ikke mulig å formulere en kort og konsis definisjon av begrepet på grunnlag av de svar som ble gitt. Det var også ulike oppfatninger om hva som er de vanligste svakhetene ved studier av trafikksikkerhetstiltak. Mange nevnte imidlertid dårlig kontroll for bakenforliggende eller andre forstyrrende variabler (”confounding factors”) som en viktig feilkilde. Når det gjaldt mulighetene for å måle kvaliteten på undersøkelser tallmessig, var de fleste ikke avvisende til tanken om dette, men det fantes en viss skepsis til om et slikt mål ville inneholde et for stort element av vilkårlighet.

En av de forskere som ble spurt, Ezra Hauer, har nylig utviklet en skala for å bedømme kvaliteten på undersøkelser som ledd i utviklingen av Highway Safety Manual i USA. Denne skalaen presenteres og visse elementer i den er benyttet i den skala som er utviklet i denne rapporten.

En pilotversjon av en skala

Det ble i 2000 utviklet en pilotversjon av en skala for kvalitet på undersøkelser om virkninger av trafikksikkerhetstiltak. Skalaen bestod av 10 poster som ble scoret med en ordinal skala. Fem forskere scoret uavhengig av hverandre fem studier om virkninger av trafikksikkerhetstiltak for å teste skalaen. Skalaen hadde en akseptabel reliabilitet. Det viste seg å være umulig å teste dens validitet. Tanken var opprinnelig å gjøre dette ved å sammenholde skalaen med en ”gullstandard”, representert ved de ledende trafikksikkerhetsforskernes oppfatning om kvalitet. Det viste seg imidlertid at disse oppfatningene var så sprikende at de ikke kunne brukes som validitetskriterium.

En annen erfaring med skalaen var at den i liten grad diskriminerte mellom de fem utvalgte undersøkelsene. Alle fikk tildelt omtrent samme poengsum, selv om det på forhånd var antatt at disse undersøkelsene representerte arbeider med ulik kvalitet. Skalaen ble forkastet og har ikke vært benyttet etter pilotstudien.

Typologi av undersøkelsesopplegg og feilkilder

Erfaringene med å gjennomgå tidligere kvalitetsskalaer, spørre ledende forskere, samt teste en pilotversjon av en skala viste at man for å utvikle en hensiktsmessig skala for kvalitet må bygge på en bred forståelse av begrepet kvalitet og en omfattende gjennomgang av faktorer som påvirker undersøkelsers kvalitet. Til dette formål er det utviklet en typologi av undersøkelsesopplegg og mulige feilkilder i hvert undersøkelsesopplegg.

De vanligste undersøkelsesopplegg i studier om virkninger av trafikksikkerhets-tiltak er (det finnes flere varianter av hvert opplegg):

1. Eksperimenter (randomiserte, kontrollerte forsøk; brukes sjelden)
2. Før-og-etter undersøkelser (mange varianter; brukes ofte)
3. Tverrsnittsstudier (uten statistisk modellering; tidligere mye brukt)
4. Case-control studier (brukes mest om skadereduserende tiltak)
5. Multivariate, statistiske ulykkesmodeller (brukes mer og mer)

6. Tidsrekkeanalyser (brukes mye om promillekjøring; lite ellers)

For hvert av disse oppleggene, ble de viktigste feilkildene knyttet til intern validitet identifisert. Med intern validitet menes grunnlaget for å trekke slutninger om årsakssammenheng mellom det undersøkte tiltaket og endringer i trafikksikkerheten.

Metodologisk forskning

For å kunne bestemme hva som skal inngå i en skala for kvalitet, må man ha kunnskap om hva som påvirker kvaliteten på en undersøkelse og hvor store virkninger ulike feilkilder kan ha på resultatene av en undersøkelse. En undersøkelse av god kvalitet kan defineres som en undersøkelse der det er lite sannsynlig at svakheter ved metoden har påvirket resultatene av undersøkelsen.

Med metodologisk forskning menes forskning der formålet er å studere hvordan ulike metodeproblemer og svakheter påvirker, eller kan påvirke, resultatene av en undersøkelse. Slik forskning kan gi et bidrag til grunnlaget for kvalitetsskalaer ved å identifisere hva som skal inngå i skalaene (hvilke sider ved metoden er relevante) og hvor stor betydning de ulike faktorene skal tillegges (betyr metodefeil A som regel mer for resultatene enn metodefeil B?)

Det ble gjort en gjennomgang av metodologisk forskning om studier av virkninger av trafikksikkerhetstiltak. Omfanget av denne forskningen varierer en god del mellom de ulike typer undersøkelsesopplegg som benyttes; mer er derfor kjent om mulige feilkilder ved noen opplegg enn ved andre. Resultatene var vanskelige å tolke. Det viste seg at selv velkjente og relativt godt utforskede feilkilder som manglende kontroll for regresjon mot gjennomsnittet slett ikke alltid påvirker resultatene av en undersøkelse nevneverdig. Når manglende kontroll for regresjon mot gjennomsnittet har betydning, viser det seg, noe overraskende, at feilen kan gå i begge retninger. Det har vært nærmest opplest og vedtatt at manglende kontroll for regresjon mot gjennomsnittet alltid og uten unntak fører til at tiltakets virkning overvurderes betydelig. Slik er det ikke. Bildet er dessverre langt mer uklart.

Gjennomgangen av metodologisk forskning ga følgelig ikke noe brukbart grunnlag for å tilordne vekter til ulike poster i en skala for kvalitet.

En skala for kvalitet på undersøkelser om virkninger av trafikksikkerhetstiltak

Til tross for at forsøkene på å etablere en forskningsmessig begrunnelse for en skala for kvalitet på undersøkelser om virkninger av trafikksikkerhetstiltak i det store og hele må betegnes som mislykkede, er en slik skala likevel foreslått i rapporten. I likhet med enhver annen skala som har vært utviklet, har denne skalaen et betydelig element av vilkårlighet. Dette synes, som rapporten viser, foreløpig å være umulig å unngå. Det valg man står overfor, er derfor enten å konkludere med at kvalitet er noe som ikke kan måles på en god nok måte, eller å måle kvalitet med en skala der de enkelte elementer ikke fullt ut kan begrunnes med henvisning til veletablert kunnskap.

Skalaen består av to deler. Den ene delen er felles for alle typer undersøkelses-opplegg. Den andre delen er skreddersydd til hver type undersøkelsesopplegg. Skalaen er normert slik at en fullkommen undersøkelse scorer 1, en helt verdiløs undersøkelse scorer 0. De ulike typene undersøkelsesopplegg er ikke innbyrdes rangordnet; en god undersøkelse kan følgelig score 1 uansett hvilket opplegg den har benyttet. Den felles delen av skalaen teller 50 %; den del som er spesifikk for hver type undersøkelsesopplegg teller 50 %. Ulike vekter er tilordnet de ulike poster som inngår i skalaen.

Skalaen bygger på kriterier for intern validitet, det vil si operasjonelle kriterier som angir hvor godt en har oppfylt betingelsene for å trekke slutninger om en årsakssammenheng mellom et tiltak og endringer i trafikksikkerheten. Disse kriteriene er utviklet og anvendt gjennom en rekke tidligere studier. Antallet poster som må sjekkes for å bedømme en undersøkelses kvalitet varierer noe mellom de ulike undersøkelsesoppleggene, og ligger mellom 10 og 20.

Skalaen er testet på 18 undersøkelser. Disse scorer verdier som ligger mellom 0.863 for den beste og 0.131 for den dårligste undersøkelsen. Skalaens reliabilitet og validitet er foreløpig ukjent og bør testes på flere undersøkelser.

Behandling av undersøkelsers kvalitet i meta-analyser

Flere tilnærmingsmåter kan tenkes til behandling av undersøkelsers kvalitet i meta-analyser. Tre tilnærmingsmåter betraktes som forsvarlige:

1. Man kan identifisere ulike aspekter ved undersøkelsers kvalitet og bruke en variabel som representerer hvert aspekt som uavhengig variabel i en meta-regresjonsanalyse.
2. Man kan tilordne hver undersøkelse en generell score for kvalitet og bruke denne som uavhengig variabel i meta-regresjonsanalyse.
3. Man kan tilordne hver undersøkelse en generell score for kvalitet mellom 0 og 1 og justere de statistiske vektene som tilordnes hver undersøkelse for undersøkelsens kvalitet. Undersøkelser som scorer nær null vil da få redusert sin vekt tilsvarende.

Alle disse tre tilnærmingsmåtene gir mening og kan forsvares. Eksempler på bruk av dem blir gitt. Tilnærmingsmåte 3 kan begrunnes med at studier av lav kvalitet kan gi mer sprikende resultater enn studier av høy kvalitet, og derfor bør telle mindre.