Oppsiktsvekkende resultater

Forskning er ikke unntatt fusk og bedrag. I 2011 ble den svært anerkjente psykologiprofessoren Stapel avslørt som en juksemaker. Dette skjedde rett før han skulle publisere nok et overraskende, men ikke utenkelig, resultat. Mennesker som spiser kjøtt er mer egoistiske enn de som er vegetarianere. Dette og andre oppsiktsvekkende resultater fra denne forskeren og hans medforfattere er nå trukket tilbake. I Norge er den største forskningsfusk-skandalen fra 2006, da det ble avslørt at kreftforskeren Sudbø hadde fabrikkert data. Fusk er likevel ikke den største trusselen mot den empiriske forskningen.

En enda større trussel, fordi den er så mye mer omfattende, er skjevheten i hva som publiseres av resultater og en (ofte ubevisst) tendens til å justere analyser og prosedyrer slik at man finner det man håper å finne. Dette fører til at en vesentlig andel av forskningsfunnen som publiseres er uriktige. John P. Ioannidis regner i artikkelen "Why most published research findings are false" (PLOS medicine) ut hvor stor andel av forskningsfunnene som er uriktige. Han finner at det i flere forskningsfelt må antas å være en svært høy andel, betydelig mer enn 50%, uriktige resultater. De fagfeltene som er verst, er de der man undersøker svært mange sammenhenger (eksplorerende undersøkelser) og har et lavt antall studieobjekter (lav statistisk styrke). Svært mange av de mest oppsiktsvekkende resultater klarer ingen uavhengige forskere å finne, sannsynligvis fordi det aldri var noe der. Uheldigvis blir resultatet fra den første undersøkelsen likevel ofte stående.

Hva så med resultater innen forskning på IT-utvikling? I forbindelse med en presentasjon på IT-konferansen International Workshop on Software Measurement (2013) framla jeg en utregning som tyder på at vi må regne med at så mye som en tredjedel av funnene fra eksperimenter på IT-utvikling er uriktige (presentasjonen kan hentes fra simula.no/publications/Simula.simula.2122/simula_pdf_file). I tillegg kommer at publiserte resultater fra IT-undersøkelser oftest viser en vesentlig større effekt enn det som reelt er tilfelle. Flere av antagelsen jeg gjorde er usikre, men det er uansett god grunn til å tro at andelen ukorrekte og oppblåste resultater innen IT-forskningen er vesentlig.

Et illustrerende eksempel på problemene med oppblåste resultater er undersøkelsene av parprogrammering. Parprogrammering er en metode der to programmerere sitter sammen og programmerer. Påstanden er at denne metoden medfører betydelig økning i kvalitet i programvaren, noe flere undersøkelser støtter opp under. Problemet er bare at det er kun småskala-undersøkelsene som viser store positive effekter. De få storskala undersøkelsene vi har viser at kvalitetseffekten av parprogrammering er relativt liten.

En bakenforliggende grunn til at småskala-undersøkelser rapporterer større effekter er at tilfeldig variasjon i prestasjoner spiller en større rolle når antall studieobjekter er lavt. Dette er den samme effekten som fører til at det er mer sannsynlig at det er stor forskjell på antall fødte gutter og jenter i løpet av en uke ved et lite enn ved et stort sykehus. Det vil med andre ord være en større mulighet for å finne store effekter i småskala enn i storskala undersøkelser. Det vi også vet er at småskala-undersøkelser som *ikke* finner noen forskjell sjeldnere får publiserer resultatene sine. Dersom en forsker på den annen side finner en effekt, til tross for å kun ha gjennomført en småskala undersøkelse, så vil småskala ofte (feilaktig!) tolkes som å styrke resultatene. Resultatet er at man får en publikasjonsskjevhet i favør av småskala studier med store effekter. I tillegg kommer at forskerens mulighet til å påvirke resultatet av en analyse er betydelig større i småskala enn i storskala undersøkelser.

Jeg gjennomførte nylig en uformell test av hvor lett det er å finne de resultatene man vil. Hypotesen jeg fremsatte var at forskere med langt navn skrev mer kompliserte tekster. Analysene fra min småskala undersøkelse ga at dette faktisk var tilfelle. Det var en høy korrelasjon (60%) mellom lengden av etternavnet og kompleksiteten (målt som "Flesh-Kincaid reading grade level") i tekstene de skrev. Jeg kunne med andre ord ha publisert et forskningsfunn, som sikkert hadde vakt oppsikt, om at forskere med langt navn skriver mer komplisert. For å finne denne sterke korrelasjonen hadde jeg ikke bedrevet forskningsfusk, men kun vært fleksibel med hensyn på hva jeg rapporterte og hvilke observasjoner jeg ekskluderte fra analysen. Jeg hadde for eksempel brukt mange typer målestørrelser for kompleksitet, men kun rapportert den som ga effekt. I tillegg hadde jeg fjernet to observasjoner, riktignok med rimelig god begrunnelse, som gjorde at resultatene ble enda mer overbevisende. Begge delene skjer regelmessig innen forskning og illustrere hvor lett det kan være å finne resultater hvis man virkelig vil.

De dypereliggende årsakene til problemene med upålitelige forskningsresultater ligger i at vitenskapelige metoder ikke brukes godt nok og dermed heller ikke klarer å forhindre at vi som forskere finner de vi ønsker å finne. Dette er det ikke lett å få bukt med så lenge det er mye lettere å få publisert oppsiktsvekkende effekter fra en småskala undersøkelse enn resultater fra undersøkelser av høyere kvalitet og større skala som viser ingen effekt. Oppmerksomheten rundt dette problemet har heldigvis økt i den senere tid. "The Economist" (October 19-25, 2013) tok for eksempel nylig opp temaet med en stor andel uriktige forskningsresultater på førstesiden med tittelen "How science goes wrong.". Det er sterkt å håpe at et økt fokus medfører forbedret forskningspraksis.

Brukerne av forskningsresultater bør være oppmerksomme på problemene med ukorrekte resultater. Særlig bør man være på vakt dersom oppsiktsvekkende resultater kommer fra småskala undersøkelser på et forskningsspørsmål ingen har undersøkt tidligere, og hvor mange ulike sammenhenger testes i samme undersøkelsen. Ofte bør man vente på uavhengige, storskala studier eller oppsummerende studier før man tar beslutninger basert på slik forskningen.

Det å finne de resultatene man ønsker å finne er ikke særegent for forskning. I de sammenhengene der det settes lavere krav til metodikk enn i forskningen, som i undersøkelser uten fagfelle-vurderinger og i erfaringsbaserte oppsummeringer, er andelen uriktige konklusjoner høyst sannsynlig enda større.