Den mesta forskningen är felaktig

Per Flensburg

Professor i informationssystem

En av världens mest citerade forskare heter John Ioannidis. Hans mest citerade verk har titeln: *Why Most Published Research Findings Are False*. Enbart denna artikel har över 10 000 citeringar på Google Scholar (Ioannidis, 2005). Han visar där med osviklig matematik att rent statistiskt stämmer hans påstående. Jag ska här försöka beskriva hans resonemang utan alltför mycket matematik. Men jag kan inte undvika att införa bokstäver som förkortningar för vissa värden.

Ioannidis arbete handlar enbart om hypotesprövningar och eftersom han är läkare hämtar han sina exempel från medicinska undersökningar, men principerna är generella för alla experimentella hypotesprövande kvantitativa undersökningar. Grundprincipen för experimentell metod är att jämföra en kontrollgrupp med en experimentgrupp, som man utsätter för någon behandling. Om experimentgruppen skiljer sig tillräckligt mycket åt från kontrollgruppen förkastar vi den så kallade nollhypotesen – att behandlingen inte haft någon effekt.

En viktig roll spelar det s k p-talet, som är sannolikheten att den verifierade hypotesen ändå inte är sann. P-talet brukar vara <0,05 och man säger att hypotesen är verifierad med 95% sannolikhet. Det betyder att i 5% av fallen är den felaktig. Hypotesen innebär att ett visst agerande, t.ex. intag av viss medicin har önskad effekt, t.ex. att patienten tillfrisknar snabbare. Man jämför då med icke-agerande, tex. intag av sockerpiller där det förväntas inte bli någon effekt.

Man kan nu göra två sorters fel: Ett typ-I-fel eller förkastningsfel är ett statistiskt fel som består av ett felaktigt förkastande av nollhypotesen. Om man drar slutsatsen att det är fel att medicinen inte har någon verkan, även om den faktiskt har det, så har man gjort ett typ I-fel. Ett typ II-fel eller acceptansfel är ett statistiskt fel som består av en felaktig acceptans av nollhypotesen. Om man drar slutsatsen att medicinen har verkan, fast den inte har det, så har man gjort ett typ II-fel.

Ioannidis tänker sig nu ett forskningsområde där man kan formulera en mängd hypoteser mellan olika variabler. En del är korrekta andra är det inte. R betecknar kvoten mellan korrekta hypoteser och alla möjliga hypoteser, helt enkelt andelen korrekta påståenden. Han förenklar nu genom att säga att det inom området finns endast en korrekt hypotes, men däremot kan man formulera många falska. Sannolikheten för att en viss hypotes ska vara korrekt är då R/(R+1). Sannolikheten att en viss undersökning ska hitta den korrekta hypotesen är 1-andel typ-II-fel och sannolikheten att undersökningen ska göra ett felaktigt påstående är 1-andel typ-1-fel. Låt oss kalla typ-I-fel för a och typ-II-fel för b. Vidare gör vi ett antal undersökningar, vi har en viss storlek, c, på vårt sample. Då kan vi ställa upp sannolikheterna i tabell 1.

Om man nu tänker sig att vi gjort en undersökning och den visar att hypotesen är korrekt så kan man undra över hur sannolikt detta är. Denna sannolikhet kallar Iaonnidis för *the positive predictive value* (PPV). Den är helt enkelt alla korrekta resultat dividerade med

samtliga resultat, dvs PPV = (1 − b)*R*/(*R* − bR + a). Lite förenklat, eftersom både a och b är små, kan man säga att det är troligare att en undersökning stämmer än att den inte gör det om (1-b)R > a. Då a brukar vara 0,05 är det alltså troligare att undersökningen är sann om (1-b)R>0,05.

Undersökning	Sant	Falskt	Totalt
Sann	c(1-b)R/(R+1)	caR/(R+1)	*c*(*R* + a – b*R*)/(*R* + 1)
Falsk	*c*b*R*/(*R* + 1)	*c*(1 − a)/(*R* + 1)	*c*(1 − a + b*R*)/(*R* + 1)
Total	cR/(R+1)	c/(R+1)	c

Tabell 1. Sannolikheter för olika utfall i en hypotesprövning

Kan man säga något om storleken på (1-b)R? b är förmodligen i samma storleksordning som a så låt oss säga 0,05. R var ju kvoten mellan alla verkligt sanna hypoteser och alla möjliga hypoteser. Vi har antagit att det finns en och endast en sann hypotes i det område vi undersöker. Så om 0,95/(antalet möjliga hypoteser) är större än 0,05 är det sannolikare att hypotesen stämmer än att den inte gör det. Antalet möjliga hypoteser får alltså inte vara större än 19.

Nu inför Ioannidis *bias*, definierat som

> ” the combination of various design, data, analysis, and presentation factors that tend to produce research findings when they should not be produced” (Ioannidis, 2005, s 697)

Exempel på bias är att vissa observationer tas bort, man testar många hypoteser och väljer ut de som verifieras etc. Ioannidis inför nu en ny variabel u, som är andelen av de resultat som på grund av bias uppfattas som korrekta. Han visar sedan att sannolikheten för att en upptäckt är korrekt minskar väsentligt med ökad bias. Ett väsentligt begrepp är *power*, vilket är sannolikheten kunna förkasta nollhypotesen om den faktiskt är falsk. Det hänger dels samman med urvalsstorlek dels med *effekt*, dvs skillnaden mellan hypotes och nollhypotes. I fig 1, som är hämtad från Ioannidis artikel visas hur sannolikheten för att en undersökning är korrekt dramatiskt minskar med ökat bias. Det verkar ju också rimligt eftersom bias innebär att man sätter slumpen ur spel och sannolikheten att resultatet är korrekt blir därmed korrupt. Vi ser också att ju större antal verkligt korrekta hypoteser det finns desto större blir påverkan från bias. I praktiken är det mycket svårt att definiera ett undersökningsområde så det finns endast en verkligt korrekt hypotes.

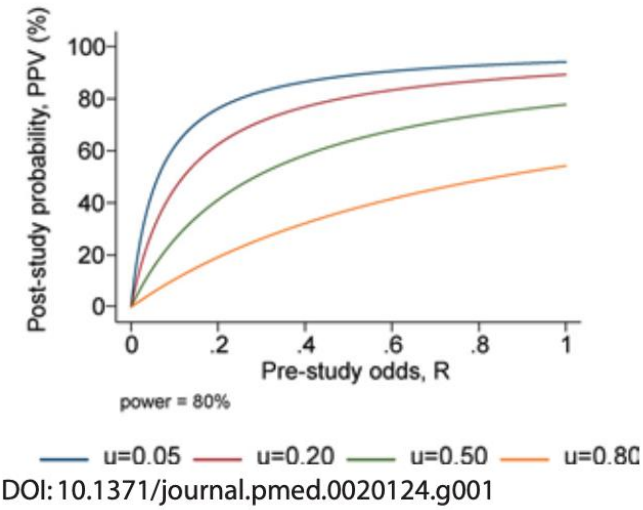

Fig 1. Hur bias påverkar PPV

Man kan nu fundera över vad som händer om det görs flera oberoende

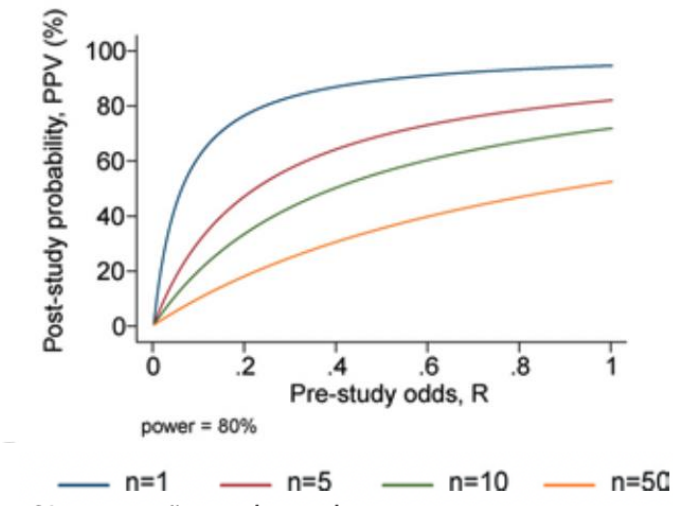

Fig 2. PPV för olika antal oberoende studier

undersökningar av samma hypotes. Det görs numera ganska ofta och man kan hitta artiklar som är sammanställning av en massa undersökningar av samma hypotes. Följaktligen ökar då sannolikheten för att någon av dem är inkorrekt och PPV minskar med antalet undersökningar. I fig 2 visar jag hur PPV förändras med antalet genomförda undersökningar utan bias. Skillnaden är här ännu mer dramatisk än vid bias.

Ioannidis formulerar nu ett antal följdsatser:

Följdsats 1: Ju färre objekt i studier som utförs inom ett vetenskapligt område, desto mindre sannolikt är forskningsresultaten korrekta

Följdsats 2: Ju mindre skillnad mellan korrekt och icke korrekt resultat är inom ett vetenskapligt område, desto mindre sannolikt är forskningsresultaten sanna.

Följdsats 3: Ju större antal och ju mindre urval av testade samband inom ett vetenskapligt område, desto mindre sannolikt är forskningsresultaten sanna.

Följdsats 4: Ju större flexibilitet i design, definitioner, resultat och analytiska lägen i ett vetenskapligt område, desto mindre sannolikt är forskningsresultaten sanna.

Följdsats 5: Ju större ekonomiska och andra intressen och fördomar som finns inom ett vetenskapligt område, desto mindre sannolikt att forskningsresultaten ska vara sanna.

Följdsats 6: Ju hetare ett vetenskapligt område (med fler vetenskapliga team inblandade), desto mindre sannolikt är forskningsresultaten sanna.

Dessa följdsatser tar upp en faktor var, men faktorerna påverkar ofta varandra. Om man tror att skillnaden mellan hypotes och nollhypotes är liten är man benägen att göra omfattande studier än där man tror skillnaden är stor.

I tabell 4 på nästa sida ser vi en simulering av olika typer av undersökningar och deras PPV. Vi ser att det krävs extremt väl designade undersökningar för att ge högt PPV. Jag känner också igen ett problem från mitt eget område, nämligen relevance versus rigour. Ska man göra en metodiskt korrekt undersökning kommer den att behandla ett isolerat och i regel tämligen ointressant och snävt område, medan intressanta undersökningar ofta kan kritiseras från metodsynpunkt.

Man kan kritisera Ioannidis från i princip två synpunkter: han ger teoretiska belägg men gör ingen empirisk undersökning som göra hans påståenden troliga och han ger ingen anvisning för hur man bedömer en undersökning utifrån hans följdsatser t.ex. Det finns förmodligen andra forskare som gjort detta. Man ska också ha klart för sig att Ioannidis undersökning gäller hypotesprövningar och då främst inom medicinen. Inom andra områden och för kvalitativa undersökningar kan man inte dra några slutsatser från hans skrift. Men det mesta av forskningen är hypotesprövande och hans undersökning är därmed ytterst väsentlig.

Table 4. PPV of Research Findings for Various Combinations of Power (1 – β), Ratio of True to Not-True Relationships *(R)*, and Bias *(u)*

1 – β	*R*	*u*	Practical Example	PPV
0.80	1:1	0.10	Adequately powered RCT with little bias and 1:1 pre-study odds	0.85
0.95	2:1	0.30	Confirmatory meta-analysis of good-quality RCTs	0.85
0.80	1:3	0.40	Meta-analysis of small inconclusive studies	0.41
0.20	1:5	0.20	Underpowered, but well-performed phase I/II RCT	0.23
0.20	1:5	0.80	Underpowered, poorly performed phase I/II RCT	0.17
0.80	1:10	0.30	Adequately powered exploratory epidemiological study	0.20
0.20	1:10	0.30	Underpowered exploratory epidemiological study	0.12
0.20	1:1,000	0.80	Discovery-oriented exploratory research with massive testing	0.0010
0.20	1:1,000	0.20	As in previous example, but with more limited bias (more standardized)	0.0015

The estimated PPVs (positive predictive values) are derived assuming $\alpha = 0.05$ for a single study.
RCT, randomized controlled trial.

Referenser

Ioannidis, J.P. (2005) 'Why most published research findings are false', *PLoS medicine*, 2(8), p. e124.