RAPPORT

Att sätta P för fusk

ADAM TAUBE, professor, institutionen för informationsvetenskap, Uppsala universitet
adam.taube@dis.uu.sea

På senare år har det dessvärre blivit aktuellt att närmare studera »fraud« i kliniska prövningar, och en speciell kommitté för detta har sett dagens ljus [1]. I ett engelsk-svenskt lexikon översätts fraud med bedrägeri, svek, svikligt förfarande. Vid kliniska prövningar har ordet kommit att gälla framför allt två speciella varianter, nämligen »data falsification« (ändring eller strykning av vissa värden) och »data fabrication« (påhittade data). Frågan är: Hur kan detta slags fiffel upptäckas i vetenskapliga manuskript eller vid granskning av grunddata från kliniska prövningar?

Det existerar faktiskt några statistiska metoder för att påvisa att data kan ha »tillverkats« på något sätt. En intressant tillämpning av sådana har presenterats i BMJ [2] där referenter fattade misstankar om ett insänt manus. De begärde därför in originaldata (som var utskrivna för hand, på papper) och utsatte dessa för närgången granskning och analys.

Manus gällde en nutritionell interventionsstudie där två grupper av patienter med koronar sjukdom fick olika diet, den ena »vanlig kost« och den andra kost med tillskott av diverse nutritionella nyttigheter. Studien påstods vara randomiserad och enkelblind, dvs deltagarna skulle ha fördelats med slumpens hjälp, och bedömarna kunde inte, åtminstone inte i utgångsläget, ha någon information om vilken grupp respektive deltagare tillhörde. Syftet var att studera om det efter två år uppstod skillnader mellan grupperna med avseende på diverse riskfaktorer för kardiovaskulär sjukdom.

Likhet i bakgrundsvariabler

Inte sällan händer det att jämförelsegrupperna i en klinisk prövning presenteras med medelvärden för ett antal relevanta bakgrundsfaktorer (»baseline characteristics«) vid starten och signifikanstest för eventuella statistiska skillnader mellan grupperna, markerade med de i medicinsk statistik så omåttligt populära P-värdena. Den (inte alltid uttalade) ambitionen är ofta att visa att grupperna är jämförbara, dvs att det inte funnits några beaktansvärda skillnader vid prövningens start.

Denna ansats har med rätta kritiserats, eftersom det även vid en korrekt randomisering mycket väl kan råka uppstå skillnader som till och med är statistiskt signifikanta. Det går därför inte att dra slutsatsen att randomiseringen »lyckats« därför att det inte föreligger några statistiskt signifikanta skillnader mellan grupperna [3].

Däremot kan det vara intressant att vända på steken på ett sätt som inte varit vanligt hitintills och – om det finns (alltför) många statistiskt signifikanta skillnader – i stället dra slutsatsen att randomiseringen inte har lyckats. Det är i den riktningen den aktuella granskningen i BMJ har bedrivits.

Det kan vara anledning att närmare beskriva vilken sannolikhetsfördelning som gäller för ett P-värde (för teoretiska detaljer se Rafe [4]). Vid ett signifikanstest för jämförelse mellan två slumpmässiga stickprov från en och samma population, så att den eventuella medelvärdesskillnaden endast är slumpens verk (dvs den s k nollhypotesen är sann), är alla värden för P i intervallet från noll till ett precis lika troliga. Med sannolikheten 0,05 kan det då i alla fall inträffa att P<0,05 (s k slumpsignifikans) (Figur 1 a).

Om det däremot är så att grupperna kommer från två olika populationer (dvs den s k nollhypotesen är inte sann) skall P-värdet förväntas ha en sned fördelning, med högre sannolikhet närmre nollpunkten, som i Figur 1 b. Då skall sannolikheten vara stor för att P<0,05.

Av detta följer att i en korrekt randomiserad, klinisk prövning med två grupper, där ett knippe bakgrundsfaktorer jämförs mellan de bägge grupperna, skall P-värdena vara rektangulärt fördelade. Om prövningen inte är korrekt randomiserad kan en starkt sned fördelning med en anhopning av låga värden förväntas.

I den misstänkta dietstudien signifikanstestades differenserna mellan de två grupperna med avseende på 21 olika bakgrundsvariabler. Det visade sig då att P-värdena för tio av dessa inte bara var <0,05, utan att de faktiskt var extremt små. Vid test av likhet mellan varianserna blev det tolv P-värden <0,05, varav flera var extremt små. Detta pekade direkt mot att data inte kunde härröra från en korrekt randomiserad studie.

Jämförelse med en annan studie

För att stärka proceduren ytterligare införskaffades data (som levererades på diskett) från en likartad studie där inga som helst misstankar om fiffel förelåg. Den gällde en mycket snarlik population patienter, med lätt hypertoni, och de två grupperna fick antingen en viss medicinsk behandling eller ingen behandling alls. Studien gick ut på att utröna om den prövade medicinen minskade risken för stroke efter två år.

Det fanns fem variabler, vilka hade mätts vid starten både i denna studie (medicinprövningen) och i den misstänkta studien (dietprövningen), nämligen kroppsvikt, kroppslängd, kolesterolvärde, diastoliskt och systoliskt blodtryck. För var och en av dessa variabler signifikanstestades dels likhet i medelvärde mellan grupperna, dels huruvida varianserna var lika eller inte. P-värdena för dessa signfikanstest återges grafiskt i Figur 2 a och 2 b.

Det befanns att i medicinprövningen förelåg ingen statistisk signifikans alls, alla de tio P-värdena blev >0,05. I dietstudien fann man att 4 av de tio P-värdena var <0,05, och det kan vara värt att notera att alla dessa gällde varianser och inte medelvärden. Detta kan vara förenligt med att man t ex bytt ut eller ändrat vissa observationsvärden för att få gruppmedelvärdena lika, men att man då kanske har råkat ändra fördelningarna så att varianserna påverkats eller så har grupperna från början varit olika ifråga om såväl medelvärde som spridning.

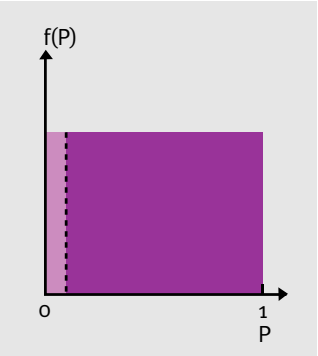
Var studien verkligen blind?

Om det vore så att dietstudien verkligen varit blind skulle ut-

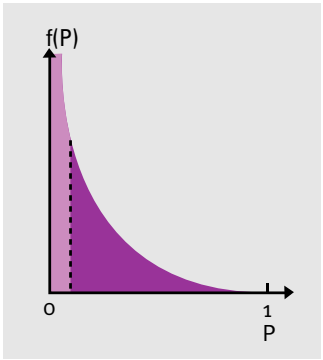
SAMMANFATTAT

Med hjälp av statistiska metoder är det möjligt att påvisa fusk i forskning. Ett framgångsrikt exempel har publicerats i BMJ, där tidskriftens referenter fattade misstankar om ett insänt manus och därför begärde in originaldata som de satte under luppen.

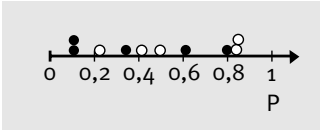
Med stor trovärdighet kunde de visa att data var manipulerade (påhittade?) och att observationsmaterialet inte kunde härröra från en korrekt randomiserad, enkelblind studie.



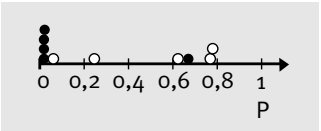
Figur 1 a. Sannolikhetsfördelning för P-värdet i ett signifikanstest, t ex jämförelse mellan två medelvärden, när den s k nollhypotesen är sann, dvs ingen reell skillnad föreligger.



Figur 1 b. Sannolikhetsfördelning för P-värdet i ett signfikanstest då nollhypotesen inte är sann.



Figur 2 a. P-värdena för signifikanstest av eventuella skillnader mellan behandlingsgrupp och kontrollgrupp med avseende på fem bakgrundsvariabler i medicinstudien. Ofyllda cirklar gäller test av medelvärden, fyllda test av varianser.



Figur 2 b. P-värden för signifikanstest av eventuella skillnader mellan gruppen med vanlig kost och gruppen med speciell kost i dietstudien för samma bakgrundsvariabler som i medicinstudien. Grafik: Helena Lunding (Fig 1 och 2)

gångsdata i de bägge randomiserade grupperna förväntas vara helt likartat behandlade. Granskarna studerade därför fördelningarna av den sista decimalen i deltagarnas värden för ett antal bakgrundsvariabler.

Det är förvisso känt (och accepterat) att den sista decimalen inte alls behöver vara jämnt fördelad över värdena 0, 1, ... 9. Det är fullt legitimt att den som mäter kan runda av på olika sätt. Blodtrycksdata visar t ex ofta en sågtandad fördelning, eftersom det är frestande att avrunda en mätning till noll eller fem. Men avrundningsmönstret bör vara detsamma i de bägge grupperna, eftersom den som gör observationerna inte skall känna till deltagarnas grupptillhörighet.

För varje variabel testades med χ^2-test med nio frihetsgrader om fördelningen av avrundningsfel var densamma i de bägge grupperna. För de 21 bakgrundsvariablerna i den misstänkta dietstudien kunde, under antagande att det endast fanns slumpmässiga skillnader mellan grupperna, förväntas drygt ett signifikant utfall med P<0,05. Det blev i verkligheten 16 P-värden <0,05. Detta resultat är förenligt med att det t ex varit två olika personer som skött datahanteringen i respektive grupp.

För de bakgrundsvariabler som fanns noterade såväl i dietstudien som i medicinstudien blev det signifikanta skillnader i avrundningmönster mellan grupperna för varje variabel i dietstudien, men inte för någon i medicinstudien.

Det kunde till slut inte råda något som helst tvivel om att dietstudien i verkligheten inte varit »enkelblind«.

Några kommentarer

Slutsatserna av granskningen blev alltså att dietstudiens observationsmaterial inte kunde härröra från en korrekt randomiserad, enkelblind studie. Om data är direkt påhittade eller om de öppet handplockats på något sätt till de två grupperna, utan randomisering, går knappast att avgöra. I vilket fall som helst kunde studien tveklöst underkännas.

I sina konklusioner hävdar granskarna att påhittade data är en sällsynt form av vetenskapligt lurendrejeri. Det är bara att hoppas att de har rätt. Vidare konstateras att de flesta vetenskapliga redogörelser publiceras utan redovisning av grundläggande data och att det på sina håll har krävts att de ursprungliga observationsvärdena bör göras tillgängliga för granskning. På den punkten finns säkert mycket mer att göra för såväl tidskriftsredaktioner som referenter.

Det konstateras vidare att de statistiska metoder som funnits för upptäckt av »fraud« endast har kommit till sparsam användning och att det ofta har påstått att det inte är möjligt, med hjälp av endast statistisk metodik, att visa att data är påhittade.

Det nya i den här relaterade granskningen är att den ger ett gott exempel på hur statistiska metoder kan tillämpas för att till och med med stor trovärdighet påvisa att data verkligen är manipulerade (påhittade?) och att vissa mönster i data är helt oförenliga med randomiseringen, speciellt i en prövning som påstås vara blind.

I praktiken är det förvånansvärt svårt att bara hitta på data, vilka sedan verkligen ser trovärdiga ut vid en mer ingående granskning. För enskilda variabler kan givetvis observationsvärden ändras så att t ex ett gruppmedelvärde blir det önskade, men sådana ändringar påverkar lätt hela fördelningens utseende.

Om det finns möjlighet, som i det här givna exemplet, att jämföra med likartade data från en annan, pålitlig studie kan just storleksordningen av varianserna ge en fingervisning om att allt inte är som det skall. Det kan säkerligen även vara mycket givande att studera korrelationsmönstret för alla noterade bakgrundsvariabler, eftersom det är svårt att bara uppfinna data, vilka i det avseendet ger ett realistiskt intryck. Hur en sådan bedömning skall gå till är inte antytt i den refererade artikeln men utgör en intressant utmaning för intresserade metodstatistiker.

Potentiella bindningar eller jävsförhållanden: Inga uppgivna.

REFERENSER

1. Buyse M, George SL, Evans S, Geller NL, Ranstam J, Scherrer B, et al. The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. Stat Med 1999;18:3435-51.
2. Al-Marzouki S, Evans S, Marshall T, Roberts I. Are these data real? Statistical methods for the detection of data fabrication in clinical trials. BMJ 2005;331:267-70.
3. Taube A. Om randomiseringens välsignelser. Läkartidningen 2000;97:4173-4.
4. Rafe MJ. A note on information seldom reported via the p value. The Americian Statistician 1999;53:303-6.