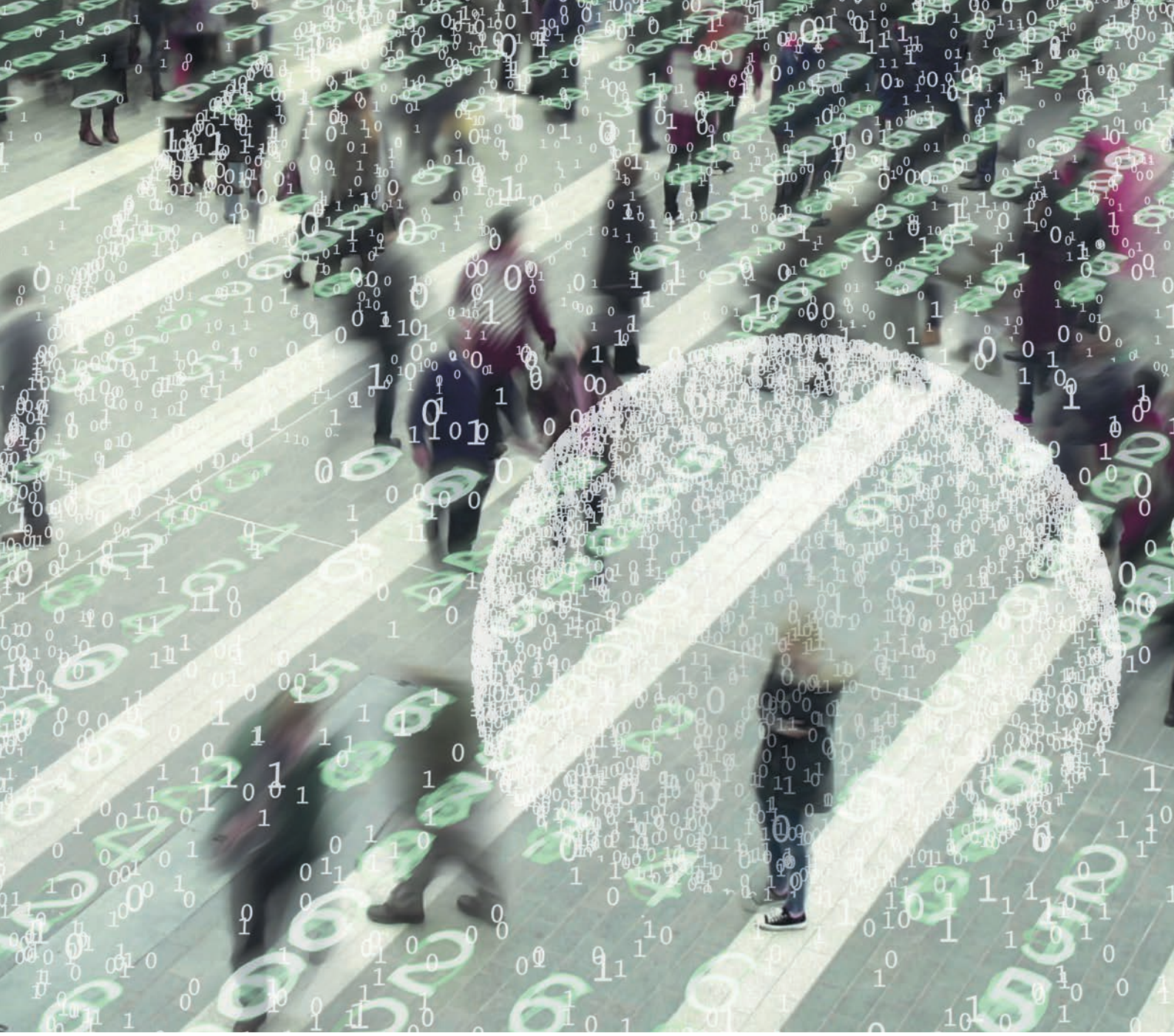
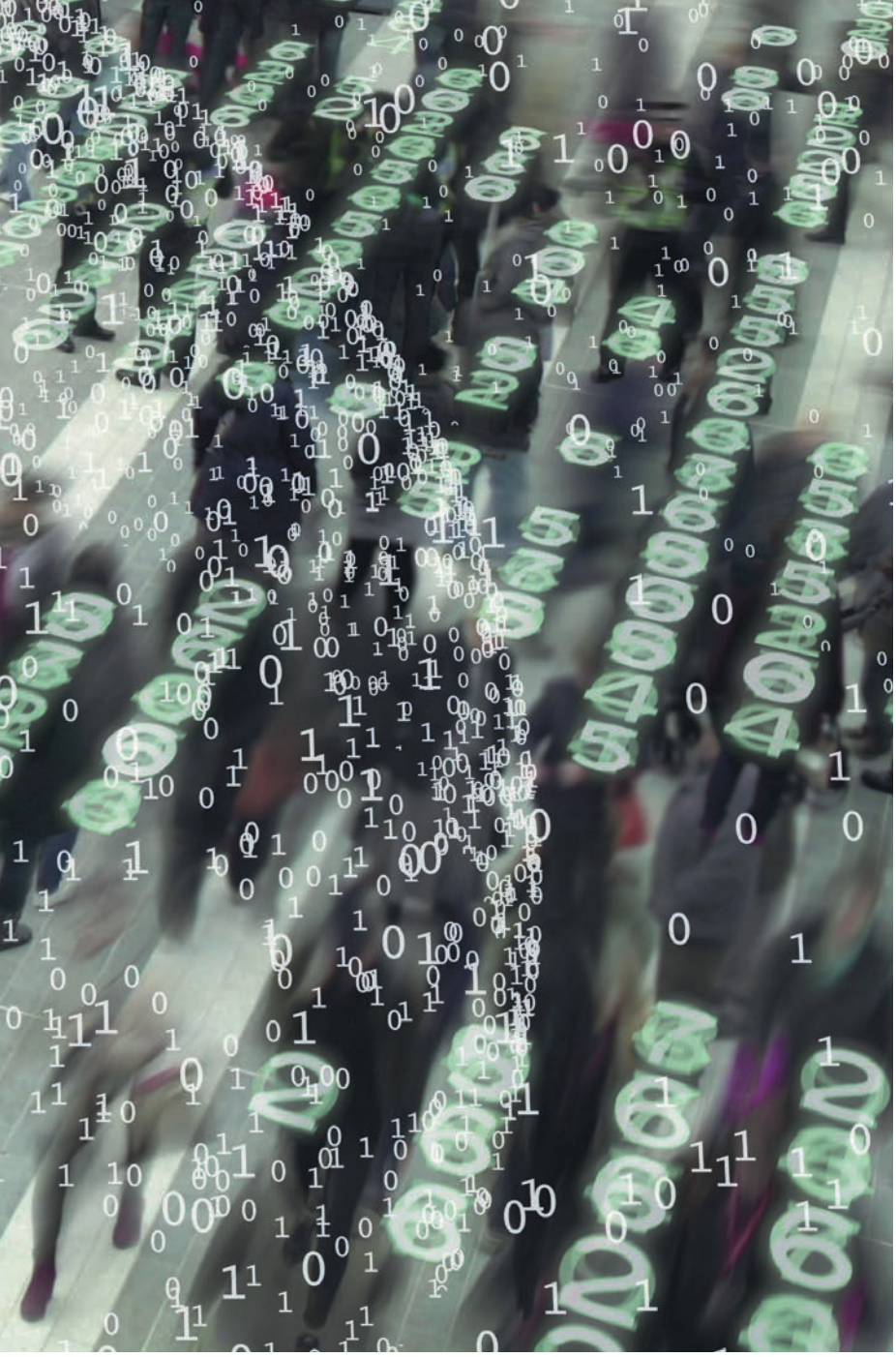


Lätt att dra fel slutsatser

Hur avancerade analyser behöver man göra vid läkemedelsstudier? Svaret är att det beror på en rad faktorer. Men klart är att det är lätt att dra fel slutsatser, beroende på hur studien är upplagd.

Om en studie på ett nytt läkemedel mot smärtor visar att 54 procent av patienterna i den grupp som fått behandling upplever smärtfrihet medan bara 39 procent i placebogruppen är smärtfria, kan man då utifrån detta dra slutsatsen att det nya läkemedlet är mer effektivt mot smärtor? Svaret är både ja och nej.

Om man bara ska uttala sig om efekterna för patienterna i den aktuella studien så kan man naturligtvis helt korrekt säga att 15 procentenheter fler patienter upplevde smärtfrihet i interventionsgruppen jämfört med placebogruppen. Men om man ska göra en klinisk studie för att kunna uttala sig om den sanna effekten av ett läkemedel, behöver man observera ett mycket stort antal patienter. De resultat man observerar i en mindre klinisk studie kan ju bero på en rad olika faktorer, bland annat slumpen. Och när man vill använda sig av resultaten i en studie för att uttala sig om totalpopulationen kommer begrepp som p-värden och konfidensintervall in i bilden. (Läs mer om dessa



p-värdet (p). Om p<0,05 så betyder det att risken för att man gör ett typ I-fel är mindre än 5 procent.

Man tar alltid en viss risk för ett typ I-fel eftersom man inte kan undersöka alla patienter med den aktuella diagnosen utan baserar slutsatsen på ett mindre urval. Eftersom gränsen för vad som är ”sant” vanligen sätts vid p<0,05 så har man samtidigt definierat hur stor risk man är villig att ta när det gäller risk för typ I-fel.

Icke-signifikans kan inte omedelbart tolkas som bevis för likhet

Ibland blir det ingen statistiskt signifikant skillnad i effekt mellan två behandlingar. Man kan då frestas att dra slutsatsen att det inte är någon skillnad i effekt och sedan belägga detta med att p-värdet är större än p>0,05. Ett stort p-värde kan bero på att nollhypotesen är sann, det vill säga att det inte är någon viktig skillnad i effekt mellan två behandlingar. En annan och minst lika vanlig orsak till stora p-värden är att studien är för liten, man har med andra ord för låg styrka (power), för att kunna visa på en signifikant skillnad i effekt.

Om man till exempel i en studie jämför två behandlingar A och B och av de 10 patienter som får A tillfrisknar 6 patienter och av de 10 patienter som får B tillfrisknar 4 patienter. Skillnaden mellan de två grupperna kan förklaras av en slump. Men om man gör samma studie på 100 + 100 patienter och resultaten

vid statistiska analyser

begrepp i Pharma Industry 1/2016 och 2/2016.)

Ibland drar man fel slutsats

Eftersom statistik, i motsatsen till matematiska bevis, aldrig bevisar att något är sant så innebär statistiska analyser att man tar en liten risk att slumpen visar en skillnad som inte är reell. Den risken kallas för typ I-fel, det vill säga att man felaktigt drar slutsatsen att det är skillnad i effekt mellan två behandlingar när det egentligen inte är det. Risken för typ I-fel mäter man med

”De resultat man observerar i en mindre klinisk studie kan ju bero på en rad olika faktorer, bland annat slumpen.

visar att 60 respektive 40 tillfrisknar så verkar det mycket mindre sannolikt att slumpen kan ha orsakat skillnaden. Om man beräknar p-värdet för de två olika studierna så blir dessa p=0,37 respektive p=0,005. Det kan tolkas som att om det egentligen inte är någon skillnad i effekt så är sannolikheten att observera en så stor skillnad 0,37 i den mindre studien och 0,005 i den större studien, det vill säga det är stor risk (eller chans) att av en slump observera skillnaden i den mindre studien men högst osannolikt att observera en så stor skillnad

i den större studien med 200 patienter. En större studie har högre styrka (power) och man kan med större säkerhet dra slutsatser om skillnader i effekt även om skillnaden inte är större än i en mindre studie.

När man felaktigt drar slutsatser om likhet, men studien egentligen har för låg power för att dra slutsatser om likhet så riskerar man att göra ett typ II-fel. Man drar en felaktig slutsats att effekten är lika men det finns egentligen skillnader som man skulle kunnat upptäcka om studien var riktigt dimensionerad.

En större studie har högre styrka (power) och man kan med större säkerhet dra slutsatser om skillnader i effekt även om skillnaden inte är större än i en mindre studie.

Finns det typ III-fel?

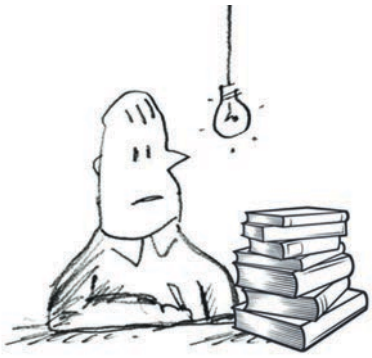
Ett vanligt fel är att man använder statistiska test och förbiser att olika villkor måste vara uppfyllda för att testet ska vara giltigt. Det kan handla om underliggande fördelning, till exempel att testet kräver att data är normalfördelade eller att modellen inte är lämpad för den frågeställning man har. Ibland kallar man det lite skämtsamt typ III-fel – allt är rätt räknat, men det blir ändå fel.

Många statistiska test kan ge ökad risk för signifikanta resultat av en slump

En vanlig felkälla är att man gör väldigt många statistiska test i en studie. Av en ren slump blir ungefär 1 av 20 test statistiskt signifikanta även om det egentligen inte finns några skillnader i effekt mellan två behandlingar. I en situation där man tillåter sig att göra väldigt många statistiska test och sedan inte korrigerar för att man gjort flera test är det stor risk att enstaka p-värden uppkommit av en slump.

I kliniska studier kontrollerar man detta genom att ange ett primärt effektmått och studien måste visa signifikans för detta effektmått för att anses lyckad. Inom epidemiologisk forskning är det inte alltid glasklart vad som var den primära frågeställningen och tolkningen av resultat därför mindre tydlig. Problematik knutet till flera statistiska test brukar man kalla multiplicitetsproblematik och det finns metoder för att korrigera p-värdena när man gör flera test samtidigt. Resultatet blir då ofta att effekterna inte anses vara statistiskt signifikanta.

I studier med flera interimsanalyser så möter man en liknande problematik – flera analystillfällen ökar risken för att något är signifikant av en slump. I studier med interimsanalyser använder man därför ibland så kallade alfa-spending functions där man endast accepterar lägre p-värden som signifikanta vid varje test.



De flesta statistiska metoder vilar på antaganden om data. Ibland kan det därför bli helt fel även om man räknat rätt. Illustration: Anders Gunè

I mindre studier är statistiska test mindre viktiga

I mindre studier är det viktigt att inte uteslutande förlita sig på statistiska test för att utvärdera data, detta eftersom man ofta har låg styrka (power) och riskerar att förbise intressanta effekter. I tidiga studier är fokus ofta säkerhet (biverkningar) och man vill också undersöka hur stora doser som kan ges utan risk för allvarliga biverkningar. Ofta ger man först en låg dos till ett litet antal, till exempel tre patienter, och så ökar man dosen för var tredje patient. Ofta kan deskriptiva metoder och grafisk utvärdering ge en bättre bild av om det finns effekter som kan vara intressanta att utvärdera i en större studie.

När det gäller utvärdering av biverkningar, även i större studier, så har man ofta låg power (styrka) för att utvärdera skillnader eftersom oftast endast ett fåtal patienter får bieffekter. Också här utvärderar man resultaten huvudsakligen med deskriptiva metoder.

Allt är inte statistik

En studie kan ha statistiskt signifikanta resultat, men ändå ha lågt bevisvärde på grund av frågetecken kring generaliserbarhet (på engelska external validity). Enkelt uttryckt handlar det om att patientpopulationen ska vara representativ för de patienter man senare önskar behandla i klinisk vardag. I en starkt selekterad patientpopulation ser man kanske tydliga effekter, men i en mer heterogen och verklighetsnära patientpopulation blir effekterna mindre tydliga.

Många möjligheter för felaktiga slutsatser

Summa summarum så finns det tyvärr många möjligheter till att dra felaktiga slutsatser. Statistisk analys handlar om att skatta effekter som man inte kan mäta exakt utan bara kan uppskatta med hjälp av kliniska studier. Ofta kan valet av statistisk metod vara avgörande och ibland verkar det enklare än det faktiskt är eftersom enkel programvara gör avancerade metoder mer tillgängliga. En bra princip är att försöka förstå underliggande antaganden och sedan försöka förenkla så mycket utan att det blir felaktigt.

ANNA TÖRNER
statistiker och verkställande direktör
Scandinavian Development Services