Bilaga 10. **Statistiska begrepp i medicinska utvärderingar**

REVIDERAD 2014

Bilagan består av två avsnitt. Det första rör de vanligaste måtten och metoderna för att bedöma validitet och tillförlitlighet i olika diagnostiska metoder. Det andra avsnittet diskuterar olika metoder för att redovisa resultat av behandlingsstudier.

Den som vill ha mer djupgående kunskaper hänvisas till läroböcker i ämnet [1–5]. Referenslistan innehåller en del litteraturtips.

Diagnostiska studier

Det finns tre grundmått: sensitivitet, specificitet och sjukdomsprevalens i den grupp som undersöks och diagnostiseras. Alla andra mått kan beräknas utifrån dessa tre. Alla mått har sina för- och nackdelar. I detta avsnitt beskriver vi närmare begreppen:

- testmetodens sensitivitet och specificitet
- prediktionsvärden
- likelihood-kvoter
- ROC-kurvor
- reliabilitetsmått.

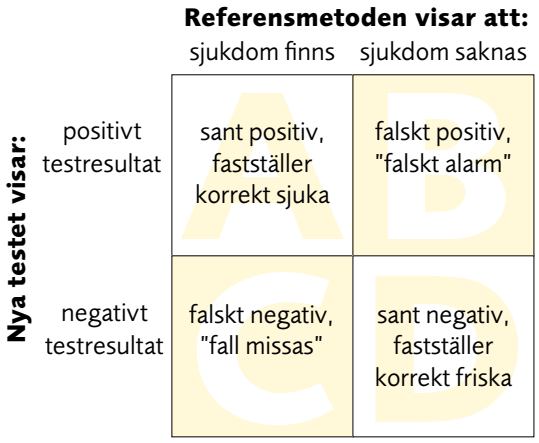
Diagnostik och riskbedömning eller prognos för att förutsäga sjukdomsutvecklingen hör nära samman eftersom diagnostik är en förutsättning för att göra en riskbedömning eller en prognos. Hur träffsäker en diagnostisk metod eller en metod för riskbedömning är, mäts också med samma mått. En bra metod ska vara tillräckligt känslig för att missa så få av dem som är/blir sjuka som möjligt och samtidigt ge så få ”falska alarm” som möjligt, det vill säga friska/icke riskindivider ska också identifieras med hög träffsäkerhet. Beräkningen förutsätter att man kan jämföra utfallet med någon standard, referensmetod eller det faktiska utfallet. Referensmetoden eller den bästa möjliga referensmetoden som internationellt kallas gold standard varierar.

När man gör en undersökning eller ett test, till exempel en kemisk analys, ett cytologiskt prov eller en röntgenundersökning, kan resultaten av tester antingen vara positiva eller negativa.

Positivt test tyder på viss sjukdom. *Negativt test* tyder på hälsa.

Med denna definition kan testresultaten beroende på om patienten verkligen är sjuk eller frisk ge fyra olika utslag:

- Sant positiva = sjuka klassificeras som sjuka (a)
- Sant negativa = friska klassificeras som friska (d)
- Falskt positiva = friska klassificeras som sjuka (b)
- Falskt negativa = sjuka klassificeras som friska (c).



Figur B10.1 Fyrfältstabell med kombinationer av testresultat och sjukdomsförekomst.

Dessa utfall brukar redovisas i en fyrfältstabell (Figur B10.1) med kombinationer av testresultat och sjukdomsförekomst.

Testmetodens sensitivitet och specificitet

Utifrån fyrfältstabellen kan testmetodens tillförlitlighet bedömas med hjälp av två mått, sensitivitet och specificitet. Naturligtvis vill man att testmetoden både ska vara *känslig*, det vill säga reagera för *alla* med sjukdomen och *specifik*, det vill säga *bara* reagera för de sjuka. Mått på hur känsligt och specifikt ett test är kallas för sensitivitet respektive specificitet (Faktaruta B10.1).

Faktaruta B10.1 Definitioner och formler för sensitivitet och specificitet.

Sensitivitet = Sannolikheten för positivt testresultat när man har sjukdomen.
Specificitet = Sannolikheten för negativt testresultat när man är frisk.

Sensitivitet = sjuka klassificerade som sjuka/alla sjuka = a/(a+c).
Specificitet = friska klassificerade som friska/alla friska = d/(b+d).

Sensitivitet och specificitet kan vardera anta värden mellan 0 och 100 procent. Ju närmare 100 procent desto bättre är det diagnostiska/prognostiska testet. Om summan av sensitivitet och specificitet är 2 är testet perfekt, det vill säga träffsäkerheten (accuracy) är 100 procent. Om sensitivitet och specificitet vardera är mindre än 0,5 (träffsäkerhet <50 %), är testet värdelöst, det vill säga det är inte bättre än slumpen.

Exempel B10.1 Räkneexempel som illustrerar måtten sensitivitet och specificitet.

	Sjuka (S+)	Friska (S–)	Summa
Positivt test (+)	950	100	1 050
Negativt test (–)	50	900	950
Totalt	1 000	1 000	2 000

Sensitivitet = 950/1 000 = 0,95 = 95 %. Specificitet = 900/1 000 = 0,90 = 90 %.

När vi ska värdera olika tester ställs vi inför en mängd beslutsproblem. Om ett test har både sensitivitet och specificitet som är högre än ett annat test, väljer man förstås det första. Oftast får vi dock göra kompromisser och välja antingen hög sensitivitet eller hög specificitet. Det gäller att se vilken typ av feldiagnos som får minst allvarliga konsekvenser.

I en del artiklar har man försökt strukturera de faktorer som har betydelse för vilka test som är lämpliga vid olika tillfällen. Ett exempel på strukturering redovisas i Exempel B10.2.

Exempel B10.2 Strukturering av faktorer som har betydelse för test.

A.	Sjukdomens prevalens (= krav på prevalens)	Förekomst av sjukdom hos den grupp som utsätts för testet
B.	"Skador" av att friska klassificeras som sjuka (= krav på specificitet)	1. Risker med att behandla friska individer 2. Ekonomiska behandlingskostnader 3. Etiska konsekvenser
C.	"Skador" av att sjuka klassificeras som friska (= krav på sensitivitet)	1. Individens risk av att ej bli behandlad 2. Befolkningens risker (smittspridning) 3. Ärftliga risker

Det är viktigt att ha ett patientperspektiv när man analyserar diagnostiska metoder. Förutom riskerna med att inte snabbt komma under behandling eller riskerna med att utsättas för felaktig behandling kan felaktiga eller osäkra besked skapa oro, ängslan eller ångest hos såväl patienter som anhöriga. Denna typ av problem är särskilt accentuerade vid screening där man undersöker en hel grupp där prevalensen av sjukdom är ganska

Exempel B10.3 Olika krav på sensitivitet och specificitet.

Onödig behandling av syfilis innebär vissa mindre risker och kostnader, men riskerna är på det hela taget små. Kraven på specificitet är därför ganska låga. Däremot är konsekvenserna av obehandlade fall allvarliga både för individen och befolkningen. Det är alltså viktigt att få tag i så många fall som möjligt. Sensitiviteten bör ofta vara hög när det gäller smittsamma sjukdomar.

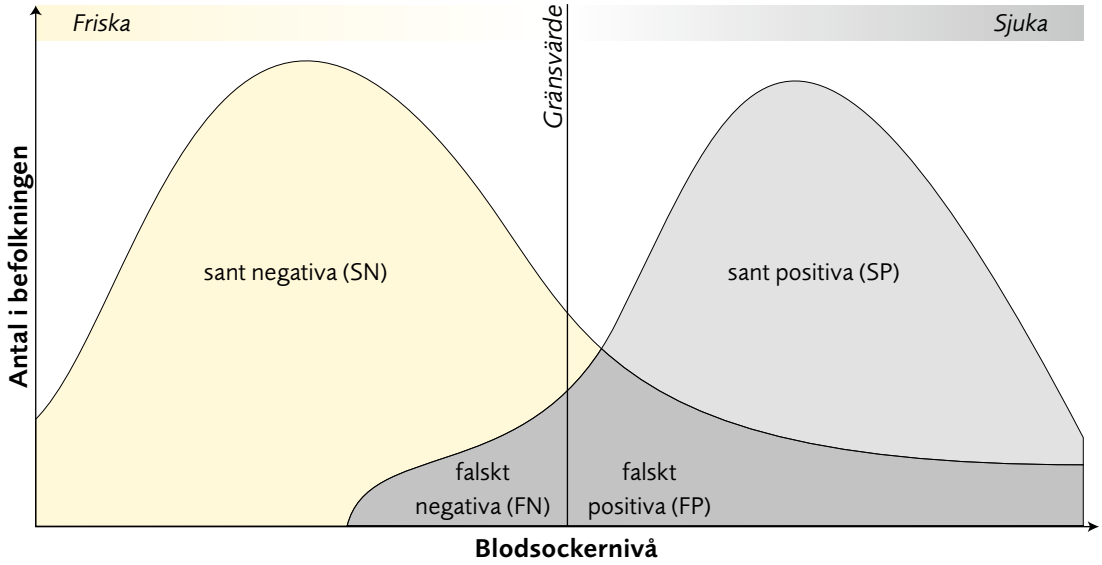
Lungcancer kan sägas vara en sjukdom med motsatt förhållande. Riskerna vid strålningsbehandling eller operation är stora. Behandlingskostnaderna är också höga. Möjligheterna att bota är för närvarande relativt dåliga och det finns inga risker för befolkningen eller för eventuella arvingar. Specificiteten i testerna bör därför vara hög. Fosterdiagnostik är ett annat exempel där kraven på specificitet måste vara mycket höga.

låg. En annan viktig aspekt av patientperspektivet på diagnostik är värdet av ett negativt test. För den enskilde individen är ett negativt testresultat på en cancerundersökning av stor betydelse för att minska oro, och därmed öka välbefinnande. Ett friande negativt test har även en ekonomisk aspekt, se Exempel B10.4. Sammanfattningsvis bör värdet av ett negativt test inte underskattas.

Exempel B10.4 Friande negativa test kan ha fördelar.

En RCT utförd i Danmark visade att 60–70 procent av de patienter som inte undersöktes med gastroskopi (= bästa möjliga referensmetod för utredning av symtom på reflux) kom tillbaka för förnyad undersökning om symtomen återkom, vilket ofta är fallet med dyspepsi [6].

Resultaten av diagnostiken uttrycks ofta i ett kvantitativt värde, till exempel blodtrycks- eller blodsockernivå. Måttligt höga blodtryck är dessutom mer en riskfaktor än en indikation på sjukdom. Gränsvärdet för blodsockernivå avgör vilka personer som antas ha diabetes. Gränsen för vad som ska betecknas som sjukt eller friskt är inte självklar. Oftast väljs en gräns, en cut-off-nivå, där man bedömer att risken för sjukdom är stor. Denna något godtyckliga gräns påverkar hur många sant och falskt positiva respektive negativa vi kommer att få (Figur B10.2). Sänks gränsvärdet för blodsockernivån kommer man att få fler falskt positiva och färre falskt negativa. Höjs gränsvärdet blir resultatet det omvända.



FN = Falskt negativa; FP = Falskt positiva; SN = Sant negativa, dvs friska klassificeras som friska; SP = Sant positiva, dvs sjuka klassificeras som sjuka

Figur B10.2 Gränsvärdet för sjukt och friskt påverkar andelen sant/falskt positiva/negativa. Figuren illustrerar fördelningen av en "frisk" och en "sjuk" befolkning där det finns en viss överlappning mellan grupperna vad gäller den mätta variabeln. Det lodrätta strecket i mitten är gränsvärdet. Om gränsvärdet förskjuts åt vänster får man fler falskt positiva och förskjuts den åt höger får man fler falskt negativa.

I praktiken baseras inte besluten på en testmetod utan oftast på en samlad bedömning av symtom och resultaten av flera diagnostiska tester. Riskerna att göra felslut är begränsade, men det kan ändå vara värt att i varje enskilt fall fundera på om sensitivitet eller specificitet är viktigast.

Sensitivitet och specificitet är två grundläggande mått som visar på testmetodens tillförlitlighet. Ett problem med dessa mått är att de förutsätter att man vet vilka som är sjuka eller friska och att jämförelse- eller referensmetoden förutsätts vara tillförlitlig. Ett annat problem är att det i en klinisk verksamhet inte säger något om sannolikheten att patienten har sjukdomen eller inte, det är bara testresultatet som är känt. Man vill gärna ha mått på möjligheterna att förutsäga (predicera) om patienten är sjuk eller frisk.

Prediktionsvärden

När man bedömer tillförlitligheten i ett test har man alltså utgått från ett facit, det vill säga man vet vilka som har en viss sjukdom eller inte. Problemet för läkarna när de ska ställa en diagnos är att de inte vet om patienten har en sjukdom eller inte. Däremot känner de till testresultatet. Det är då mer intressant att veta hur sannolikt det är att patienten har sjukdomen, när testresultatet är positivt. Detta kallar vi *positivt prediktionsvärde* och är en så kallad betingad sannolikhet, det vill säga sannolikheten för att de som testas har sjukdomen betingat av att testresultatet är positivt.

Ett sätt att beräkna det positiva prediktionsvärdet är att sätta alla sant positiva resultat i relation till alla positiva testresultat, det vill säga enligt exemplet nedan:

Positivt prediktionsvärde (PPV) = a/a+b = 950 sant positiva/1 050 positiva testresultat = 0,905 = 90,5 %.

På motsvarande sätt är man intresserad av att veta om patienten inte har sjukdomen om testresultatet är negativt. Detta kallas *negativt prediktionsvärde* och är sannolikheten för att de som testas inte har sjukdomen betingat av att testresultatet är negativt, det vill säga man dividerar sant negativa med alla negativa testresultat. Enligt exemplet nedan får vi:

Negativt prediktionsvärde (NPV) = d/c+d = 900 sant negativa/950 negativa testresultat = 0,947 = 94,7 procent.

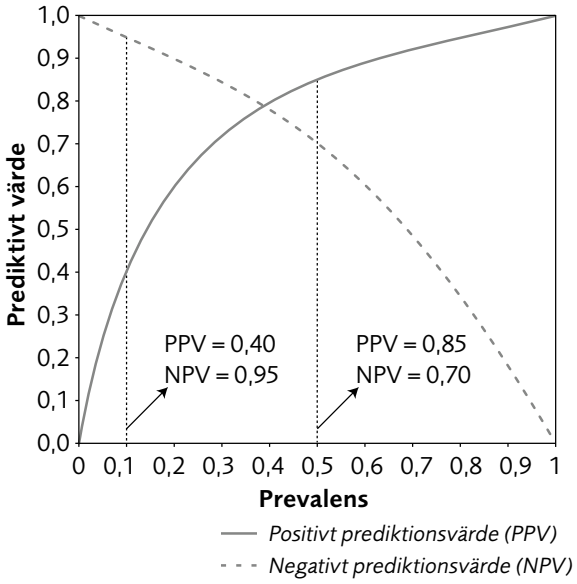
Tre faktorer påverkar prediktionsvärdena. Självklart påverkar testmetodens sensitivitet och specificitet möjligheterna att förutsäga om patienten är sjuk eller frisk. Det är inte lika intuitivt självklart att prevalensen i den grupp som undersöks påverkar prediktionsvärdena. Att så dock är fallet kan visas genom att vi ändrar prevalensen i vårt hypotetiska exempel, se Exempel B10.5.

Exempel B10.5 Positiva och negativa prediktionsvärden.

I Exempel B10.1 hade hälften av de undersökta sjukdomen i fråga, det vill säga en prevalens på 50 procent. Om vi har samma sensitivitet och specificitet, men ändrar sjukdomsprevalensen i den grupp som undersökts till 5 procent blir prediktionsvärdena helt annorlunda.

Antal	Sjuka (S+)	Friska (S–)	Summa
Positivt test (+)	95	190	285
Negativt test (–)	5	1 710	1 715
Totalt	100	1 900	2 000

I detta fall med en prevalens på 5 procent blir det positiva prediktionsvärdet (PPV) = 95/285 = 33,3 procent jämfört med 90,5 procent när prevalensen var 50 procent. Det negativa prediktionsvärdet (NPV) blir 99,7 procent (1 710/1 715) jämfört med 94,7 procent när prevalensen var 50 procent. Slutsatsen är att det positiva prediktionsvärdet sjunker om prevalensen är låg medan det negativa prediktionsvärdet blir högre om prevalensen är låg.



Figur B10.3 Diagram som illustrerar hur det positiva och negativa prediktionsvärdet påverkas av sjukdomsprevalensen. I exemplet är sensitiviteten (andelen sanna positiva) 0,6 och specificiteten (andelen sanna negativa) 0,9. Vid en sjukdomsprevalens på 10 procent blir det positiva prediktiva värdet (PPV) 0,40 medan det negativa prediktiva värdet (NPV) är 0,95. Om sjukdomsprevalensen är 50 procent, ökar PPV till 0,85, medan NPV sjunker till 0,70.

Möjligheterna att med en viss testmetod förutsäga om en patient är sjuk eller inte varierar alltså beroende på i vilken situation metoden används. Vid screening där de flesta som testas är friska, det vill säga en låg prevalens, blir det positiva prediktionsvärdet relativt lågt. Om man testar metoden på en grupp patienter som remitterats pga stark klinisk misstanke om sjukdom kan prevalensen i den gruppen vara mycket hög och man får därigenom ett högt prediktionsvärde. En distriktsläkare träffar kanske bara på sjukdomen i ett fall per 10 000 patienter medan en specialist på universitetssjukhuset får vissa patientkategorier där nästan var femte patient kan ha sjukdomen ifråga. Ett test på en universitetsklinik med selekterade patienter, det vill säga hög prevalens, fungerar kanske inte lika bra i primärvården där sjukdomsförekomsten är låg. I den kliniska situationen bör man alltså alltid fundera på i vilken situation man befinner sig och bedöma hur prevalensen kan påverka beslutet.

Likelihood-kvoter

Det kan vara en fördel att ha ett mått på ett tests prestanda, ett mått som sammanfattar sensitivitet och specificitet och som är oberoende av sjukdomsprevalensen. Likelihood-kvoter (LHR) är sådana mått som uttrycks i termer av odds. Ett odds är sannolikheten för att en viss händelse ska inträffa dividerad med sannolikheten att den inte ska inträffa. Oddskvoter för likelihood-kvoter redovisas som en sannolikhet för ett visst testresultat om man är sjuk dividerat med sannolikheten för samma testresultat om man är frisk.

Eftersom testresultaten antingen kan vara positiva eller negativa kan man få två likelihood-kvoter.

En *positiv likelihood-kvot (LHR+)* beskriver sannolikheten att vara testpositiv om man är sjuk dividerat med sannolikheten att vara testpositiv om man är frisk = *sensitivitet/(1–specificitet).* Det kan också uttryckas som oddset för att ett positivt test kommer från en person med sjukdomen istället för en utan den. Ju högre LHR+-värde desto högre är sannolikheten att personen har sjukdomen i fråga.

En *negativ likelihood-kvot (LHR–)* beskriver sannolikheten att vara testnegativ om man är sjuk dividerat med sannolikheten att vara testnegativ om man är frisk = *(1–sensitivitet)/specificitet.* Det kan också uttryckas som oddset för ett negativt test kommer från en med sjukdomen istället för en utan den. Ju lägre LHR–-värde desto mindre är sannolikheten att personen har sjukdomen i fråga.

Exempel B10.6 Beräkning av likelihood-kvoter.

Utifrån vårt tidigare fiktiva exempel kan likelihood-kvoterna beräknas enligt följande:

Antal	Sjuka (S+)	Friska (S–)	Summa
Positivt test (+)	950	100	1 050
Negativt test (–)	50	900	950
Totalt	1 000	1 000	2 000

LHR+ = sensitivitet/(1–specificitet) = 0,95/(1–0,90) = 9,5.
Oddset för att ett positivt test kommer från en sjuk person istället för en frisk är 9,5.

LHR– = (1–sensitivitet)/specificitet = (1–0,95)/0,90 = 0,055.
Oddset för att ett negativt test kommer från en sjuk person istället för en frisk är 0,055.

Det är inte helt självklart hur likelihood-kvoter ska tolkas. Grunden är dock att ju högre oddset är för positiva likelihood-kvoter (LHR+) desto bättre är testet på att fastställa sjukdom. När det gäller negativa likelihood-kvoter (LHR–) är låga odds att föredra eftersom det minskar sannolikheten att personen har sjukdomen. Det finns dock inga klara gränser för vad som är ett stort odds för att personen har sjukdomen. Enkla tumregler har föreslagits, se Faktaruta B10.2.

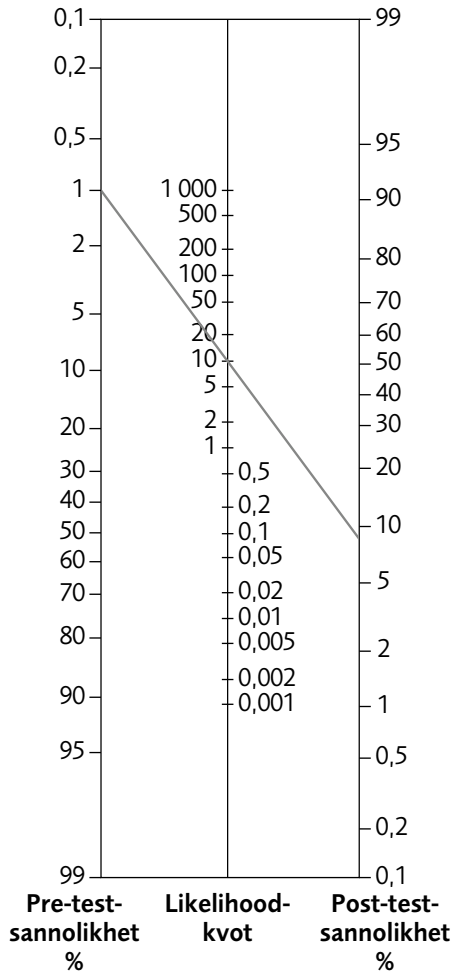
Faktaruta B10.2 Tumregler för likelihood-kvoter.

LHR+	Sannolikhet att personen har sjukdomen
>10	Stor eller mycket stor ökning
5–10	Måttlig ökning
2–5	Liten ökning, kan vara betydelsefull
1–2	Mycket liten ökning, sällan betydelsefull
LHR–	**Sannolikhet att personen har sjukdomen**
<0,1	Mycket stor minskning
0,1–0,2	Måttlig minskning
0,2–0,5	Liten minskning, kan vara betydelsefull
0,5–1	Mycket liten minskning, sällan betydelsefull

LHR+ = Positiv likelihood-kvot; LHR– = Negativ likelihood-kvot

Eftersom diagnostiken oftast bygger på en samlad värdering av symtom, anamnes och flera olika tester kan det vara värdefullt att titta på odds före (pre-test) och odds efter (post-test), det vill säga före och efter att man fått resultatet av ett test. Matematiskt kan detta uttryckas som produkten av odds före och LHR, det vill säga *odds efter = odds före x LHR*. Oddset före bildas som prevalensen/(1–prevalensen) och kan antingen baseras på faktiska data om prevalensen i den grupp som studeras eller på subjektiva uppskattningar om risken att patienten i den grupp som studeras har en viss sjukdom.

Oddset kan också redovisas som eftersannolikhet (post-test probability) med formeln *efter-sannolikhet = oddset efter/(oddset efter + 1)*. Observera att oddset efter positivt test är lika med det positiva prediktionsvärdet. För att beräkna oddset efter kan man använda beräkningsprogram eller avläsa det via nomogram (Figur B10.4).



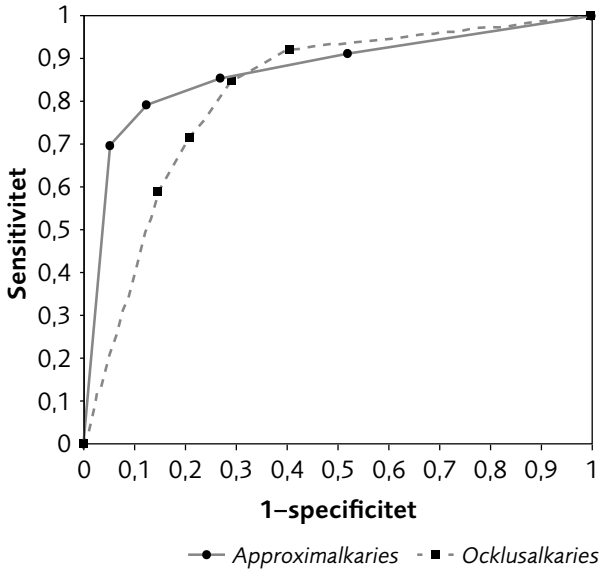
Figur B10.4 Nomogram.

Exempel B10.7 Odds före och efter resultat av test.

I vårt fiktiva exempel med positiva testresultat blir oddset före = 0,5/(1–0,5) = 1. Oddset efter = oddset före x LHR+ = 1 x 9,5 = 9,5. Efter-sannolikhet = oddset efter/(oddset efter + 1) = 9,5/(9,5 + 1) = 90,5 procent.

ROC(receiver operating characteristics)-kurvor

Receiver operating characteristics (ROC) är en grafisk redovisning av testmetoders prestanda. Den anger sambandet mellan en testmetods sensitivitet och andel falskt positiva (1–specificitet). Genom att plotta sensitivitet gentemot 1–specificitet för varje cut-off (tröskelvärde) och sedan förbinda punkterna får man en så kallad ROC-kurva. ROC-analys är hämtad från forskning om mottagning av radio- och radarsignaler, där ett signal–brus-förhållande analyseras. Fördelen med ROC-analys är att man kan analysera en metods prestanda när man har flera möjliga nivåer för gränsdragning mellan ”sjukt” och ”friskt”. I Figur B10.5 ges ett exempel på hur ROC kan användas. Rent matematiskt är den optimala gränsdragningen mellan sjukt och friskt i den punkt som ligger närmast diagrammets övre vänstra hörn. Det bästa valet i en klinisk situation behöver dock inte vara där. Beroende på situationen, kan man välja att prioritera en stor andel sant positiva och acceptera att andelen falskt positiva blir relativt stor, eller välja en relativt låg andel sant positiva för att undvika en hög andel falskt positiva. Se tidigare diskussion. Om man jämför två testmetoder där den ena ligger helt ovanför den andra så är den förra helt klart bäst.



Figur B10.5 ROC-kurvor.
Receiver operating characteristic (ROC) med sanna positiva värden (sensitivitet) på Y-axeln och falska positiva värden (1–specificitet) på X-axeln. Kurvorna visar sambandet mellan sant positiva värden och falskt positiva värden för dentinkaries (djup karies innanför emaljen) på approximalytor (tandytor mot varandra) och ocklusalytor (tuggytor). Exemplet är från en in vitro-studie med extraherade tänder (bästa möjliga referensmetod = histologisk verifikation), där undersökare fick ange graden av säkerhet för dentinkaries på röntgenbilden på en skala från 0="säkert ingen dentinkaries" till 5="helt säkert dentinkaries" [7].

Arean under ROC-kurvan är ett sammantaget mått på testets prestanda och är lika med sannolikheten att en slumpmässigt vald person med sjukdomen har ett högre värde än en slumpmässigt vald person utan sjukdomen. Arean 1 eller 100 procent anger att det är ett perfekt test och mindre än 0,5 eller 50 procent att det inte är bättre än slumpen.

Reliabilitetsmått

Reliabiliteten (tillförlitligheten, precisionen) hos en diagnostisk metod är ett uttryck för hur väl till exempel en kariesdiagnos överensstämmer mellan olika undersökare eller hur väl samma undersökare kan upprepa en specifik diagnos vid ett senare tillfälle. Överensstämmelsen inom eller mellan undersökare kan sammanfattas på olika sätt; som hur många procent av ett antal diagnoser man är överens om, som korrelationskoefficienter eller som kappavärde. Kappavärdet används ofta för att beskriva en diagnostisk metods reliabilitet. Kappavärdet är den observerade överensstämmelsen justerad för sannolikheten att överensstämmelsen beror på slumpen. I Exempel B10.8 redovisas ett exempel på hur kappavärdet beräknas.

Exempel B10.8 Beräkning av kappavärde: ett hypotetiskt exempel.

Röntgenbilder av 60 tandytor (approximalytor) tolkas av två undersökare med avseende på förekomst av djup karies (dentinkaries). Positiv = karies; negativ = ingen karies.

a) Resultatet av upprepad tolkning av röntgenbilder av 60 tandytor gav följande resultat:

1:a tolkningen	2:a tolkningen		
	Positiv	Negativ	Totalt
Positiv	11	11	22
Negativ	3	35	38
Totalt	14	46	60

Vid första tolkningen hade 22 ytor karies och vid andra tolkningen som gjordes ”blint” (ovetande om vad observatören kommit fram till första gången) hade 14 ytor karies. Elva ytor var positiva båda gångerna och 35 ytor var negativa vid båda tillfällena. Den procentuella överensstämmelsen var alltså (11+35)/60=0,77 eller 77 procent. Detta mått på observatörsvariation är dock missledande, eftersom man ignorerar att de två tolkningarna kan ge samma resultat beroende på slumpen.

En del av överensstämmelsen hade även slumpmässigt kunna inträffa, det vill säga att båda tolkningarna gett positiva respektive negativa utfall.

Exemplet fortsätter på nästa sida

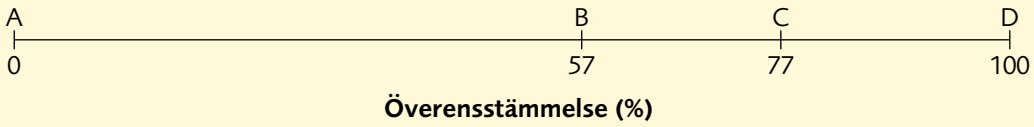
Exempel B10.8 fortsättning

Det är enkelt att beräkna det förväntade slumpmässiga resultatet om diagnoserna var oberoende av varandra. Om man tänker sig att observatören andra gången slumpmässigt väljer 14 röntgenbilder och benämner dem ”positiva”. Man kan då förvänta sig att 14x22/60=5,13 bilder kommer att bli ”positiva” vid båda tillfällena och att 46x38/60=29,13 kommer att bli ”negativa” båda gångerna. Det förväntade resultatet beroende på slumpen är illustrerat i b). Man kan alltså beräkna att det förväntade resultatet som beror på slumpen är (5,13+29,13)/60=0,57 eller 57 procent. När man tar detta i betraktande, blir den observerade procentuella överensstämmelsen på 77 procent mindre imponerande.

b) Det förväntade slumpmässiga resultatet om diagnoserna var oberoende av varandra.

1:a tolkningen	**2:a tolkningen**		
	Positiv	**Negativ**	**Totalt**
Positiv	5,13	16,87	22,00
Negativ	8,87	29,13	38,00
Totalt	14,00	46,00	60,00

Det är detta problem som formaliserats i den så kallade kappastatistiken, som relaterar den observerade överensstämmelsen till den överensstämmelse som kan bero på slumpen. I det beskrivna exemplet är kappavärdet 47 procent, vilket betyder att skillnaden mellan den observerade överensstämmelsen och den slumpberoende överensstämmelsen (77–57 %) är endast 47 procent av skillnaden mellan perfekt överensstämmelse och slumpmässig överensstämmelse (100–57 %). Detta kan illustreras med en figur:



Kappa = (C–B)/(D–B) = (77–57)/(100–57) = 0,47 eller 47%.

A = Total brist på överensstämmelse; B = Den förväntade överensstämmelsen pga slumpen; C = Den observerade överensstämmelsen; D = Perfekt överensstämmelse

Tumregler för att värdera kappavärdet har utvecklats, se Faktaruta B10.3. Det finns även andra mått för att bedöma överensstämmelsen som inte redovisas här.

Faktaruta B10.3 Tumregler för att värdera kappavärdet.

Kappavärde	Grad av överensstämmelse
≤0,20	Dålig
0,21–0,40	Svag
0,41–0,60	Måttlig
0,61–0,80	Bra
0,81–1,00	Mycket bra

Övergripande diskussion om olika mätmetoders för- och nackdelar

Den här översikten visar att det finns många sätt att mäta diagnostiska testmetoders värde. Alla har sina för- och nackdelar. I vissa situationer är det viktigast att ha en hög sensitivitet, i andra att ha en hög specificitet. Sensitivitet och specificitet mäter testmetodens tillförlitlighet, men säger inte hur säkert man kan veta om patienterna som testats har sjukdomen eller ej. Det beror bland annat på prevalensen av sjukdomen i den grupp som undersöks. I sådana sammanhang är positiva och negativa prediktionsvärden att föredra eftersom de tar hänsyn till såväl sensitivitet, specificitet som prevalens. Likelihood-kvoter uttrycks ofta som odds och har fördelen att man sammanfattar sensitivitet och specificitet i ett mått som är oberoende av prevalensen. Å andra sidan påverkar alltid prevalensen i den grupp som studeras hur väl en specifik testmetod fungerar i praktiken. Att mäta oddset före och efter en testmetod kan vara ett bra sätt för att bedöma den marginella nyttan av att ta ytterligare ett test. De grundläggande måtten är dock prevalens, sensitivitet och specificitet. Med hjälp av dessa tre grundmått kan alla andra mått beräknas.

Behandlingsstudier

Grundläggande statistiska mått när det gäller att bedöma utfallet av olika typer av behandlingsinsatser är:

- absoluta och relativa risker
- risk- och oddskvoter
- nödvändigt antal behandlingar för att förebygga ogynnsamma händelser (NNT)
- konfidensintervall och hantering av slumpen.

Effekten av en behandling kan uttryckas på många olika sätt och i såväl absoluta som relativa tal. Det finns för- och nackdelar med de flesta och man ska vara medveten om att

vi tolkar data olika beroende på hur effekten uttrycks. Det finns alltså en risk att de som redovisar resultaten av en studie väljer mått som passar med det intryck de vill förmedla. Förutom grundläggande mått redovisas också hur man kan göra metaanalyser för att få en mer samlad bild av resultaten.

De statistiska mått som används och redovisas nedan är desamma för samtliga studietyper. Behovet av att kontrollera för bakomliggande faktorer är dock mycket större för kohort- och fall–kontrollstudier. Ett sätt att statistiskt hantera sådana metodproblem är att stratifiera analyserna i olika undergrupper eller göra regressionsanalyser.

Utgångspunkten för beräkning av olika statistiska mått är en fyrfältstabell där vi på raderna har de två behandlingsinsatser eller metoder som jämförs, det vill säga en försöks- och en kontrollgrupp (Exempel B10.9). I kolumnerna anges antalet personer och antal utfall/händelser i respektive grupp. Utfallet kan vara dödsfall, sjukdomsincidens etc. Med denna tabell kan olika riskmått beräknas.

Absoluta och relativa risker

Riskmått kan uttryckas som absolut risk och relativ risk för en händelse eller ett utfall som är förknippat med åtgärden/interventionen i fråga. Vid en jämförelse mellan försöks- och kontrollgrupp talar man om absolut och relativ riskreduktion. Utifrån fyrfältstabellen kan vi definiera begreppen och illustrera beräkningarna med ett fiktivt exempel, se Exempel B10.9.

Exempel B10.9 Absoluta och relativa risker.

Studiegrupper	Antal personer	Antal utfall/händelser
Försöksgrupp, t ex läkemedel A	a) 144	c) 19
Kontrollgrupp, t ex placebo, annat läkemedel	b) 148	d) 24

Absolut risk i försöksgruppen = c/a = 19/144 = 0,132 = 13,2 procent.

Absolut risk i kontrollgruppen = d/b = 24/148 = 0,162 = 16,2 procent.

Absolut riskreduktion = absolut risk i kontrollgruppen – absolut risk i försöksgruppen = 16,2 – 13,2 = 3,0 procentenheter.

Relativ riskreduktion = absolut riskreduktion/absolut risk i kontrollgruppen = 3,0/16,2 = 18,5 procent.

Fördelen med ett absolut riskmått är att det anger den absoluta risken för att en viss händelse ska inträffa utan att jämföra risken med någon annan. Eftersom det lyckligtvis är så att negativa utfall är relativt sällsynta blir riskreduktionen ofta väldigt låg. Den relativa riskreduktionen visar effekten i relation till ett alternativ och blir därför högre. Av förklarliga skäl föredrar industrin som har kommersiella intressen oftast att redovisa resultaten i relativa termer eftersom det förstärker de positiva effekterna. Effekten upplevs som större.

Risk- och oddskvoter

De två grundläggande måtten när man jämför två åtgärder eller behandlingsalternativ är relativ risk (RR) och oddskvot (odds ratio, OR). Definition och räkneexempel presenteras i Faktaruta B10.4 och Exempel B10.10.

Faktaruta B10.4 Formler för relativ risk och oddskvot.

Relativ risk (RR) = absolut risk i försöksgruppen/absolut risk i kontrollgruppen.

Odds = sannolikhet att ha utfallet (t ex avlida)/sannolikhet att inte ha utfallet (t ex leva).

Oddskvot (OR) = odds i försöksgruppen/odds i kontrollgruppen.

Exempel B10.10 Risk- och oddskvoter.

Studiegrupper	Antal personer	Antal utfall/ händelser	Risk	Odds
Försöksgrupp, t ex läkemedel A	a) 144	c) 19	19/144=0,132	19/(144–19)=0,152
Kontrollgrupp, t ex placebo, annat läkemedel	b) 148	d) 24	24/148=0,162	24/(148–24)=0,194

Relativ risk (RR) = 0,132/0,162 = 0,81.

Oddskvot (OR) = 0,152/0,194 = 0,78.

R_K = grundrisk i kontrollgruppen = 0,162

För de flesta är det svårare att tolka en oddskvot än en riskkvot. Skillnaden mellan odds och risk blir mindre ju mer sällsynt händelsen är. Om oddset är 1 mot 10 eller 0,1 blir risken 0,91 (1 av 11). När händelserna är vanliga blir skillnaden mellan odds och risk stor. En risk på 0,5 är detsamma som ett odds som är 1 och en risk på 0,95 är detsamma som ett odds på 19.

Genom multiplikation med 100 omformas riskkvoter (men inte oddskvoter) till procenttal. Ofta uttrycker man effekten i termer av riskminskning. Om riskkvoten är 0,25 blir riskminskningen 75 procent, det vill säga 100 x (1–0,25). Den kliniska betydelsen av en riskminskning kan ses vid jämförelse med den absoluta risken i kontrollgruppen. Om händelsen drabbar 80 procent i kontrollgruppen och 60 procent i experimentgruppen kan det ha en helt annan innebörd än om händelsen drabbar 4 procent i kontrollgruppen och 3 procent i experimentgruppen.

Det blir problem om en oddskvot tolkas som en riskkvot. Vid åtgärder som ökar sannolikheten för händelser kommer oddskvoten att bli större än riskkvoten, särskilt när händelsen är vanlig, varför feltolkning leder till överskattning av behandlingseffekten. För åtgärder som minskar sannolikheten för händelser kommer oddskvoten att bli mindre än riskkvoten, så även i denna situation är det lätt att misstolka resultatet, vilket är ganska vanligt.

Oddskvoter kan omvandlas till riskkvoten och riskkvoter till oddskvoter. Detta sker genom jämförelse med risken i kontrollgruppen (R_K) (eller genom något annat mått på grundrisk). Det görs på följande sätt:

$$RR = \frac{OR}{1-R_K(1-OR)} \quad \text{medför} \quad OR = \frac{RR(1-R_K)}{1-(R_K x RR)}$$

Dessa omvandlingar kan också behövas när man i olika studier ömsevis använt det ena eller andra måttet.

Nödvändiga antal behandlingar för att förebygga ogynnsamma händelser (NNT)

För att få en intuitiv uppfattning om möjligheterna att hjälpa patienter brukar man beräkna det antal patienter som måste behandlas för att förebygga en ogynnsam händelse, till exempel en hjärtinfarkt, död eller återfall i cancer. Detta mått kallas för nödvändiga antal behandlingar (number needed to treat, NNT) och beräknas som det inverterade värdet av den absoluta riskreduktionen. Om den absoluta riskreduktionen är 3 procent

blir NNT=1/0,03~34, det vill säga i genomsnitt krävs det att man behandlar 34 patienter för att undvika en ogynnsam händelse som död.

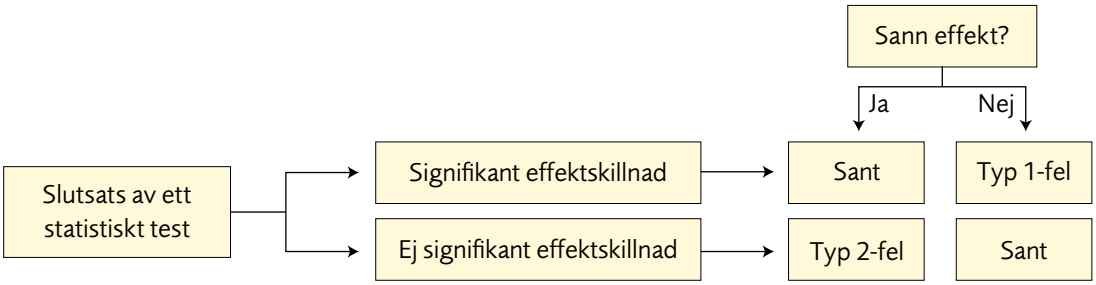
Exempel B10.11 NNT för att undvika en stroke.

I en metaanalys av sju studier där man jämförde betablockadsbehandling av mild till måttlig hypertoni med placebo var den absoluta riskreduktionen i stroke 0,22 procent [8]. Det innebär att 455 patienter (1/0,0022) behöver behandlas under 3–5 år för att undvika en stroke.

När det gäller risker går det att beräkna ett motsvarande mått, det vill säga nödvändiga antal behandlingar för att en skada ska uppkomma (number needed to harm, NNH).

Konfidensintervall och hantering av slumpen

Slumpen kan göra att de resultat vi finner inte är sanna. Det finns två sätt att göra fel med en statistisk metod. Antingen kan man dra en falskt positiv slutsats, det vill säga att en ny behandlingsmetod är effektiv när den inte är det. Denna typ av felaktig slutsats kallas Typ 1-fel eller α-fel och definieras som sannolikheten att man ser en skillnad i behandlingseffekt när det inte finns någon (Figur B10.6). Den andra typen av fel är när man drar slutsatsen att den nya behandlingsmetoden inte är effektivare fast den i verkligheten är det. Man drar då en falskt negativ slutsats vilket kallas Typ 2-fel eller β-fel.

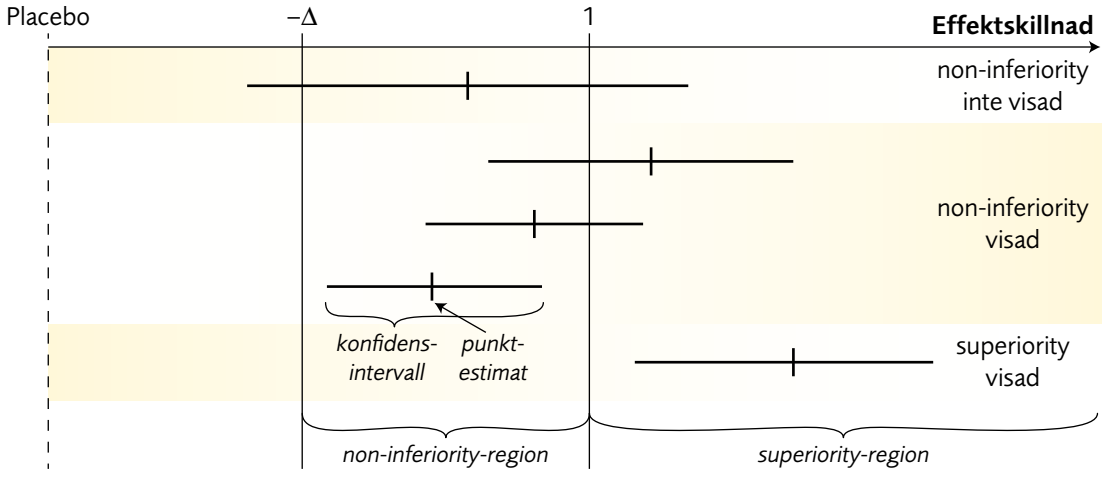


Figur B10.6 Samband mellan resultaten av ett statistiskt test och den sanna effekten mellan två behandlingsinsatser.

Konfidensintervallen visar på den statistiska osäkerheten i det urval som studeras. De bygger oftast på att man accepterar att Typ 1-fel (α) är mindre än 0,05. Små urval ger stora konfidensintervall. För beräkning av konfidensintervall hänvisas till statistiska metodböcker [1–3]. Vanligtvis väljs 95-procentiga eller 99-procentiga konfidensintervall. Med ett 95-procentigt konfidensintervall kommer det skattade värdet i 95 av 100 fall att

hamna inom konfidensintervallet. Det är viktigt att inse att vi talar om sannolikhetsbedömningar och inte om en sanning. Även om resultaten inte är statistiskt signifikanta så kan det finnas ett verkligt samband.

Resultaten kan presenteras som i Figur B10.7. På engelska kallas det för en forest plot där mittlinjen 1 indikerar att de två interventionerna (åtgärderna) är likvärdiga. Om hela konfidensintervallet för oddskvoten eller den relativa risken ligger under 1 är den studerade interventionen statistiskt säkerställt bättre. Ligger konfidensintervallet för OR/RR helt över 1 är interventionen statistiskt säkerställt sämre. Korsar konfidensintervallet linjen 1 har vi ingen statistiskt säkerställd skillnad. Punkten i mitten anger det uppskattade värdet (punktestimatet).



Figur B10.7 Punktestimat och konfidensintervall för olika utfall av effektskillnader mellan kontrollbehandling (=1) och experimentbehandling i en RCT. Konfidensintervallets nedre gräns avgör om resultatet är förenligt med superiority respektive non-inferiority.

Superiority versus non-inferiority-studier

Studier som visar klart bättre effekt för en behandling (överlägsenhet, superiority) har ett inbyggt kvalitetsmått om randomiseringen är korrekt och man kan utesluta bias (brister i blindning). Brister i studiens kvalitet i övrigt kan möjligen göra att eventuella effektskillnader mellan behandlingsarmarna underskattas. En förutsättning är dock att behandlingen i kontrollgruppen är optimal, exempelvis att dosval i jämförelsegrupperna i läkemedelsprövningar varit rättvist.

Ibland är det dock lämpligt att designa en studie för att visa att det inte föreligger någon (kliniskt relevant) skillnad (non-inferiority). Till exempel är det i vissa situationer oetiskt att använda placebokontroll samtidigt som man ibland inte kan förvänta sig att ett läke-

medel visar bättre effekt än standardbehandling. Om ett läkemedel/metod har en klart bättre säkerhetsprofil än standardbehandling men sannolikt inte är effektivare är en non-inferiority-design nödvändig. När målet att visa en behandlings överlägsenhet inte uppnås, kan resultatet bedömas som non-inferior om prövningens kvalitet tillåter detta [9].

Vid granskning av non-inferiority-studier är det viktigt att bedöma vad som är en rimlig kliniskt relevant effektskillnad och underlaget för prövarnas uppskattning av detta förspecificerade delta (Δ). Valet av Δ måste uppfylla två krav:

1. Det krävs att Δ är specificerat så att man kan vara rimligt övertygad om att Δ klart skiljer sig från placebo beaktande osäkerhet rörande förändringar i övriga faktorer som över tid förbättrat prognosen av tillståndet och därmed påverkar effekten. Som ett riktmärke är det rimligt att kräva att minst 50 procent av referensbehandlingens effekt kvarstår (Figur B10.7).

2. Valet av Δ ska återspegla en rimlig klinisk uppfattning av vad som är relevant effektskillnad.

Vid beräkning av Δ enligt punkt 1 ovan är det viktigt att man utgår från en eller helst flera RCT som jämfört kontrollbehandlingen med placebo (alternativt annan kontroll vid hårda utfallsvariabler) under samma betingelser (patientgrupp, dos, stadium, utfallsvariabel, responskriterier, etc) som den aktuella studien. Dessa bör heller inte vara alltför gamla eftersom nya undersökningsmetoder, samtidig behandling och kontrollbehandlingens relevans påverkas av utvecklingen.

Det prespecificerade Δ är ett planeringsinstrument för att dimensionera studien, den finala non-inferiority-värderingen är en nytta/risk-värdering av det observerade utfallet vid vilken man inte är bunden vid det *a piori* uppsatta Δ. Vad som kan anses utgöra en rimlig klinisk relevans (se nedan) är således inte bara en effektfråga utan är också avhängig respektive behandlings säkerhetsprofil. En klar skillnad i biverkningsmönster till fördel för experimentarmen kan göra det rimligt att acceptera ett större Δ.

Vidare är bedömning av datakvalitet, studiens upplägg och genomförande fundamental vid granskningen av en non-inferiority-studie. Brister på inom dessa områden (patientpopulation/diagnostik, mätmetoder, mätfrekvens, val av dos, omhändertagande- och andra studieeffekter) tenderar alla att försämra möjligheten att upptäcka verkliga skillnader mellan behandlingar och därmed ”simulera” likhet. Även stora bortfall gynnar likhet varför såväl per protokollanalys som ITT-analys också krävs för att non-inferiority ska accepteras [10].

Klinisk relevans

Klinisk relevans är ett lika viktigt som svårdefinierat begrepp. En kliniskt relevant effekt är den effekt som är meningsfull att upptäcka alternativt utesluta, beroende på studiedesign. Begreppet är grundläggande i planeringen av jämförande studier (power i superiority-studier och deltaberäkning i non-inferiority-studier). Även storleken på fas II-studier bestäms utifrån antagande om vad som är en kliniskt relevant effekt vilken är värd att upptäcka. Dessutom är det krav vid godkännande av ett nytt läkemedel att man kunnat demonstrera en kliniskt relevant effekt. Vid värderingen av de observerade resultaten ska man inte fästa sig vid vad som specificerats vid beräkning av den nödvändiga studiestorleken. Denna beräkning går ut på att få en hög sannolikhet (power) för ett statistiskt signifikant resultat givet att den specificerade skillnaden är sann. Den garanterar inte hög sannolikhet för att observera en skillnad som är minst lika stor som den specificerade. Sannolikheten för detta är bara 50 procent givet att antagandet är sant. Med hög power kommer även mindre observerade skillnader att bli statistiskt säkerställda och det är dessa som ska bedömas vägt mot observerade och inte observerade potentiella biverkningar för att avgöra om resultatet är kliniskt relevant.

Vad som är en kliniskt relevant effekt beror i första hand på vems perspektiv man utgår ifrån; patientens, anhörigas, behandlande läkare, myndighet eller third party payers. Åtskilliga undersökningar har kunnat slå fast att vad som uppfattas som en relevant effekt i relation till risker och kostnader kraftigt påverkas av om man är föremål för åtgärden, ordinatör eller bara bystander.

Exempel B10.12 Behandlingars värde och klinisk relevans.

I en studie av attityder till kemoterapi vid behandling av cancer tillfrågades patienter, friska kontroller, specialistläkare, allmänläkare och sköterskor om vilket behandlingsalternativ de skulle välja i olika situationer. Resultatet visade att patienterna var mycket mer riskbenägna och villiga att acceptera biverkningar än övriga grupperna, trots låga odds för tillfrisknande [11]. Den här typen av information är av stor vikt vid diskussioner om behandlingars värde och klinisk relevans.

Naturligtvis varierar uppfattningen kraftigt även mellan individer inom en grupp och sannolikt också från ett tillfälle till ett annat och mellan olika perioder i livet hos samma individ. Exempelvis har sjuka individers förmåga att hantera sin sjukdom stor betydelse för förändringar över tid när det gäller uppfattningen om klinisk relevans. Det är dock rimligt att anta att för en enskild individ uppfattas en behandling med stor effekt som mer relevant än en behandling med liten effekt, allt annat lika.

En effekt kan i praktiken vara mer eller mindre relevant (kontinuerlig variabel). Oftast hanterar man dock klinisk relevans som en dikotom variabel, det vill säga man önskar fastställa vad som är en relevant respektive icke relevant klinisk effekt. Detta är en praktisk förenkling av verkligheten på samma sätt som frågan om vem som kan anses vara responder eller ej.

Medan statistisk signifikans är lätt att definiera, närmast av axiomatisk natur och ett begrepp som det skrivits tusentals vetenskapliga publikationer om, är litteratur omkring begreppet klinisk relevans sparsam. Klinisk relevans går inte att slå fast utifrån kvantitativa metoder även om den till syvende och sist måste anges med ett kvantitativt mått. Det är snarare kvalitativa data som kan bibringa en uppfattning om vad som är en kliniskt relevant effekt. Dessa blir dock alltid föremål för en subjektiv bedömning vilken ånyo är avhängig bedömarens perspektiv, det vill säga vilken grupp man tillhör.

Patientens upplevelser kan i allmänhet fångas på tre nivåer; i form av patient satisfaction, som hälsorelaterad livskvalitet (HRQoL) eller i form av kvalitativ forskning, till exempel intervjuer.

- Det är viktigt att vara tydlig med skillnad mellan registrering av patient satisfaction vilket är ett behandlingsutfall som kan vara kopplat till HRQoL (men inte nödvändigtvis behöver vara det) och HRQoL.

- HRQoL är index som mer övergripande påverkas av individens hela livssituation, det som betecknas som individens livsvärld, och som i sig inrymmer psyklogiska, sociala och medicinska aspekter som av många anledningar fluktuerar över tid för alla individer.

- Forskningsintervjuer är ett mänskligt samtal med syfte att få en beskrivning av den intervjuade för att beskriva och/eller tolka upplevelser/innebörd eller mening i förhållande till den intervjuades livsvärld. Intervjuerna genomförs som ett samtal omkring upplevelserna där den intervjuade fritt berättar och där intervjuarens uppgift är att ställa stödfrågor om så behövs.

För att fånga patientens upplevelse av den kliniska relevansen behövs sannolikt alla tre nivåerna. Troligen kommer dock den mest relevanta informationen fram i forskningsintervjun.

Vad som uppfattas som en kliniskt relevant effekt är också avhängigt en behandlings säkerhetsprofil. Effekten av en behandling med små risker kan uppfattas som kliniskt relevant medan samma effekt hos en behandling med uttalade risker kan uppfattas

som varande icke relevant. Med andra ord borde man snarare tala om kliniskt relevant ”nytta/risk-balans” snarare än kliniskt relevant effekt.

När det rör sig om symtomatisk behandling av benigna sjukdomar där otillräcklig behandling inte försvårar prognosen på sikt är det rimligt att lämna bedömningen om storleken av effekt i relation till frekvens av biverkningar av toleranskaraktär till patienten. Om det däremot finns allvarligare säkerhetsproblem, empiriskt identifierade eller potentiella utifrån prekliniska data, verkningsmekanism, klasseffekter etcetera, är det i många fall rimligt att nytta/risk-värderingen görs av någon annan än patienten.

Vid allvarliga sjukdomar där det troliga utfallet är död är förhållandet det motsatta. Frekvent intolerans som påverkar livskvalitet blir viktig, särskilt om effektfördelen rör sig om små skillnader i tid till död i sjukdomen. Däremot är allvarliga infrekventa biverkningar mindre viktiga i denna situation, i varje fall om risken för dem är klart mindre än risken för död i sjukdomen.

Klinisk relevans ska inte blandas ihop med kostnadseffektivitet även om det är ett grundvillkor för att en metod ska kunna anses kostnadseffektiv att dess effekt bedömts vara kliniskt relevant, medan det omvända inte gäller.

Den politiska dimensionen handlar om vem (och var) som ska bedöma vad som är en kliniskt relevant effekt. Fördelen med att låta den välinformerade patienten i varje situation själv eller i samråd med sin behandlare bedöma vad som är kliniskt relevant är att den ovan beskrivna inter- och intraindividuella variabiliteten då har hanterats. Nackdelen är att rättighetsaspekter (jämlik vård) inte tillgodosetts alls. Omvänt är fördelen med en myndighetsbedömning att rättviseaspekten och frågan om optimalt resursutnyttjande möjligen hanterats bättre. Dessutom har inte alla patienter förmåga att ta till sig information så att de kan anses vara välinformerade. Därmed måste frågan om klinisk relevans i många fall ändå hanteras av behandlare, högre administrativ nivå eller av myndighet. Det är i denna situation viktigt att beslutsfattaren är väl införstådd med målgruppens uppfattning om klinisk relevans.

Referenser

1. Altman DG. Practical statistics for medical research. Chapman and Hall/Crc Pree Llc; 1991.

2. Bland M. An introduction to medical statistics. Oxford University Press, 3rd ed; 2000.

3. Bring J, Taube A. Introduktion till medicinsk statistik. Studentlitteratur; 2006.

4. Taube A, Malmquist J. Räkna med vad du tror. Bayes' sats i diagnostiken. Läkartidningen 2001;98:2910-3.

5. Taube A, Malmquist J. Räkna med vad du tror. Bayes – inte P-värdet – mäter tilltron. Läkartidningen 2001;98:3208-11.

6. Bytzer P, Hansen JM, Schaffalitzky de Muckadell OB. Empirical H2-blocker therapy or prompt endoscopy in management of dyspepsia. Lancet 1984; 343:811-6.

7. Espelid I, Tveit AB. A comparison of radiographic occlusal and approximal caries diagnoses made by 240 dentists. Acta Odontol Scand 2001;59:285-9.

8. Lindholm LH, Carlberg B, Samuelsson O. Should β blockers remain first choice in the treatment of primary hypertension? A meta-anlysis. Lancet 2005;366:1545-53.

9. Guideline on the choice of the non-inferiority margin. http://www.ema.europa.eu/pdfs/human/ewp/215899en.pdf

10. Points to consider on switching between superiority and non-inferiority. http://www.ema.europa.eu/pdfs/human/ewp/048299en.pdf

11. Slevin ML, Stubbs L, Plant HJ, Wilson P, Gregory WM, Armes PJ, Downer SM. Attitudes to chemotherapy: comparing views of patients with cancer with those of doctors, nurses, and general public. BMJ 1990;300:1458-60.