

# Metoder ved ækvivalens- og noninferioritetsundersøgelser

Overlæge Erik Christensen

Bispebjerg Hospital, Intern Medicinsk Klinik I

Den randomiserede kliniske undersøgelse (*randomized clinical trial* (RCT)) anses for at være den bedste metode til sammenligning af behandlings effekt [1]. En RCT er hyppigst designet til at vise, at en ny behandling er bedre end en etableret behandling eller placebo, dvs. den er planlagt og udført som en superioritetsundersøgelse. Somme tider er formålet med en RCT blot at vise, at en ny behandling er lige så god eller i det mindste ikke dårligere end en etableret behandling, dvs. den er planlagt og udført som en ækvivalensundersøgelse eller noninferioritetsundersøgelse. Da de forskellige typer RCT har forskelligt formål, afviger de fra hinanden med hensyn til metodologiske forhold. Denne artikel gennemgår kortfattet de vigtigste metodologiske forskelle. I denne sammenhæng vil de vigtigste relevante statistiske begreber blive beskrevet. Der henvises til [2] for en detaljeret gennemgang med eksempler på beregning af materialestørrelse og konfidensintervaller.

Oftentimes tages en manglende signifikant forskel mellem behandlingerne i en superioritetsundersøgelse til indtægt for, at behandlingerne har den samme effekt eller er ækvivalente. En sådan konklusion er ikke korrekt, fordi der er en betydelig risiko for at overse en klinisk relevant forskel pga. et utilstrækkeligt antal patienter.

Et vigtigt aspekt i planlægningen er at estimere det nødvendige antal patienter [3]. De forskellige undersøgelsestyper varierer på dette punkt. Normalt vil en undersøgelse have til formål at påvise, at en ny behandling er bedre end en etableret behandling eller placebo (superioritetsundersøgelse). For at kunne estimere det nødvendige antal patienter er det nødvendigt at overveje følgende:

1. Hvor meget skal den nye behandling være bedre end referencebehandlingen? Denne ekstra effekt af den nye behandling i forhold til den gamle behandling kaldes den mindste relevante forskel eller den kliniske signifikans. Den betegnes ofte med det græske bogstav  $\Delta$  (delta).

2. Hvor meget vil forskellen i effekt imellem de to grupper blive påvirket af tilfældige faktorer? Som enhver anden biologisk måling er behandlingseffekter genstand for en vis tilfældig variation, som man må estimere og tage hensyn til. Størrelsen af variationen kan beskrives statistisk som standardafvigelsen  $S$  eller variansen  $S^2$ . Variansen af effektvariablen må bestemmes i et pilotstudie eller på basis af tidligere publicerede undersøgelser.

Undersøgelsen bør designes til så præcist som muligt at påvise den sande forskel i effekt mellem behandlingerne. Pga. tilfældig variation vil det endelige resultat af en undersøgelse afvige i større eller mindre grad fra den sande forskel. Hvis f.eks. nulhypotesen  $H_0$  (ingen forskel) var sand, kunne undersøgelsen i visse tilfælde alligevel vise en forskel i effekt (falsk positivt resultat). Denne type fejl, som kaldes type 1-fejl, ville have den konsekvens, at en ineffektiv behandling ville blive introduceret. Hvis på den anden side den alternative hypotese  $H_{\Delta}$  at forskellen i effekt var  $\Delta$  var sand, kunne undersøgelsen i visse tilfælde falde således ud, at en behandlingsforskel ikke ville kunne påvises (falsk negativt resultat). Denne type fejl, der kaldes type 2-fejl, ville have som konsekvens, at en effektiv behandling ville blive forkastet. Ideelt burde både type 1- og type 2-fejlsrisici være tæt ved nul, men dette ville kræve meget store undersøgelser. Begrænsede resurser og patientantal gør det nødvendigt at acceptere visse mindre risici for type 1- og type 2-fejl.

Hyppigt vil risikoen for type 1-fejl  $\alpha$  blive fastlagt til 5%. Risikoen for type 2-fejl  $\beta$  vil hyppigt blive fastlagt til 10-20%. Da en given værdi af  $\Delta$  altid er enten over eller under nul (svarende til  $H_0$ ), vil risikoen for type 2-fejl  $\beta$  altid være ensidig. Jo mindre  $\beta$ , jo større komplementær sandsynlighed  $1-\beta$  for at

## Faktaboks

Et negativt resultat i en superioritetsundersøgelse beviser ikke, at de undersøgte behandlinger er ækvivalente. Ækvivalens- og noninferioritetsundersøgelser kræver et betydeligt større antal patienter end en tilsvarende superioritetsundersøgelse gør.

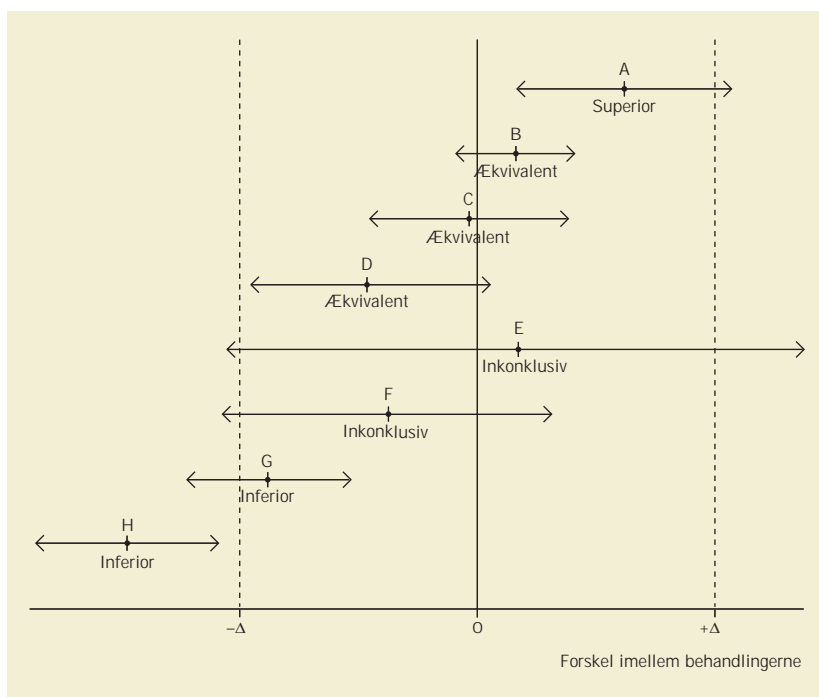
Ækvivalensmarginen  $\Delta$  skal være mindre end den mindste klinisk betydningsfulde forskel.

Ækvivalens er kun påvist, hvis sikkerhedsgrænserne for den påviste forskel ligger helt inden for intervallet  $-\Delta$  til  $+\Delta$ . Noninferioritet er kun påvist, hvis den nedre sikkerhedsgrænse for forskellen (nye behandling – kontrolbehandling) ligger over  $-\Delta$ .

Ækvivalens- og noninferioritetsundersøgelser skal udføres på samme måde som superioritetsundersøgelser (selektionskriterier, effektvariabel, behandlingsskema, blinding, skjult randomisering, followup mv.). Ækvivalens eller noninferioritet påvises kun, hvis både *intention to treat*- og per protokol-resultatet viser dette.

## VIDENSKAB OG PRAKSIS | STATUSARTIKEL

**Figur 1.** Eksempler på observerede effektforskelle (ny behandling – kontrolbehandling) med 95%-sikkerhedsintervaller og konklusioner. A: Den nye behandling er signifikant bedre end kontrolbehandlingen, men forskellens størrelse kan være klinisk betydningsløs. B-D: Behandlingerne kan antages at være ækvivalente. E-F: Resultatet er inkonklusivt. G: Den nye behandling er signifikant dårligere end kontrolbehandlingen, men forskellens størrelse kan være klinisk betydningsløs. H: Den nye behandling er signifikant dårligere end kontrolbehandlingen.



acceptere  $H_{\text{delta}}$  når denne faktisk er sand.  $1-\beta$  kaldes styrken af undersøgelsen, fordi den angiver sandsynligheden for at påvise  $\Delta$ , hvis denne forskel faktisk eksisterer.

### Ækvivalensundersøgelser

Formålet med en ækvivalensundersøgelse er at påvise identiske effekter af de to behandlinger, der sammenlignes [4]. En ækvivalensundersøgelse vil være relevant, hvis den nye behandling er enklere, billigere eller behæftet med færre bivirkninger, selv om den ikke forventes at have en større terapeutisk effekt end kontrolbehandlingen.

Komplet ækvivalens betyder en  $\Delta$ -værdi på nul. Da de formler, der benyttes til beregning af antallet af patienter, har  $\Delta$  i nævneren, kan beregningen ikke gennemføres for  $\Delta=0$ . Det ville svare til, at man skulle benytte et uendeligt stort antal patienter. Division med en meget lille  $\Delta$  vil tilsvarende resultere i et urealistisk stort antal patienter. Som et anvendeligt kompromis er målet for en ækvivalensundersøgelse derfor at påvise, at forskellen i effekt imellem to behandlinger ligger inden for et fastlagt lille interval fra  $-\Delta$  til  $+\Delta$ .

Det er afgørende vigtigt at specificere en relevant størrelse af  $\Delta$  [5]. Dette er ikke enkelt. Man skal have som formål at reducere chancen så meget som muligt for, at den nye behandling accepteres, hvis den faktisk er dårligere end kontrolbehandlingen. Derfor skal  $\Delta$  vælges ganske lille og i alle tilfælde mindre end den mindste værdi, der repræsenterer en klinisk betydningsfuld forskel. Som en grov generel vejledende regel skal  $\Delta$  være mindre end halvdelen af den værdi,

man ville benytte i en superioritetsundersøgelse. En lille  $\Delta$  vil betyde en betydelig forøgelse af det nødvendige antal patienter. Ækvivalens imellem behandlingerne vil være påvist, hvis konfidensintervallet for forskellen i effekt mellem behandlingerne viser sig at ligge fuldstændig inden for intervallet  $-\Delta$  til  $+\Delta$ . **Figur 1** illustrerer de konklusioner, som kan drages ud fra beliggenheden af sikkerhedsgrænserne for forskellen i effekt påvist i undersøgelsen.

I ækvivalensundersøgelsen er rollerne for nulhypotese og alternativ hypotese byttet om. I ækvivalensundersøgelse er den relevante nulhypotese, at der er en forskel på mindst  $\Delta$ , og målet for undersøgelsen er at forkaste dette til fordel for den alternative hypotese, at der ikke er nogen forskel. Til trods for dette er metoden til materialestørrelsesberegning den samme som i superioritetsundersøgelsen.

### Noninferioritetsundersøgelser

Noninferioritetsundersøgelsen, som er i familie med ækvivalensundersøgelsen, har ikke til formål at vise ækvivalens, men kun at vise at den nye behandling ikke er dårligere end referencebehandlingen.

Noninferioritetsundersøgelsen er således designet til at påvise, at forskellen i effekt (ny behandling-kontrolbehandling) ikke må være mindre end  $-\Delta$ . Noninferioritet af den nye behandling vil således være påvist, hvis den nedre sikkerhedsgrænse for forskellen i effekt mellem de to behandlinger viser sig at ligge over  $-\Delta$ . Positionen af den øvre sikkerhedsgrænse er ikke af primær interesse. Noninferioritetsundersøgelsen er designet som en ensidig undersøgelse. Af denne grund er det

## VIDENSKAB OG PRAKSIS | STATUSARTIKEL

nødvendige antal patienter mindre end i en tilsvarende ækvivalensundersøgelse.

### Andre vigtige forskelle

Da formålet med en ækvivalens- eller noninferioritetsundersøgelse er at påvise ækvivalens mellem behandlingerne eller noninferioritet af den nye behandling, er der ikke den samme motivation som i en superioritetsundersøgelse for at fjerne faktorer, der kan formindske forskellen mellem behandlingerne. Nogle tilfælde af påvist ækvivalens vil derfor kunne tilskrives mangler ved undersøgelsen, såsom lille materialestørrelse, manglende dobbeltblindning, manglende skjult tilfældig allokering, ukorrekt dosering af medikamenter, indflydelse af samtidig givet anden medicin eller spontan helbredelse af patienterne uden medicinsk intervention.

En ækvivalens- eller noninferioritetsundersøgelse skal så tæt som muligt afspejle de metoder, man har brugt i tidligere superioritetsundersøgelser, som har testet effekten af kontrolbehandlingen. Specielt er det vigtigt, at inklusions- og eksklusionskriterierne, som definerer patientpopulationen, blindingen, randomiseringen, doseringsskemaet for standardbehandlingen, brugen af anden medicin og andre interventioner, den primære responsvariabel og dennes registrering er de samme som i de forudgående superioritetsundersøgelser, der har evalueret referencebehandlingen. Samtidig skal man være opmærksom på patientkomplians, respons i en indkørsperiode og omfanget af patienttab og årsagerne hertil. Disse faktorer må ikke afvige fra tidligere superioritetsundersøgelser.

Et vigtigt punkt i analysen vedrører, hvorvidt man skal benytte »*intention-to-treat*« eller »*per protokol-analyse*«. I en superioritetsundersøgelse, hvor formålet er at undersøge, om to behandlinger har forskellig effekt, vil intention to treat-analysen almindeligvis være konservativ, idet inklusion af protokolsvigtende patienter og udgåede patienter sædvanligvis vil have en tendens til at gøre resultaterne for de to behandlinger mere ens. For en ækvivalens- eller noninferioritetsundersøgelse vil en sådan effekt ikke længere være konservativ, idet enhver sløring af forskellen mellem behandlingsgrupperne vil øge chancen for at finde ækvivalens eller noninferioritet.

En per protokol-analyse sammenligner resultaterne svarende til den behandling, patienterne rent faktisk fik, og inkluderer kun de patienter, som opfyldte inklusionskriterierne og fulgte protokollen fuldstændigt. I en superioritetsundersøgelse vil denne type analyse have en tendens til at øge forskellen mellem behandlingerne snarere end at formindske den, fordi noninformativ støj fjernes. I en ækvivalens- eller noninferioritetsundersøgelse skal begge analysetyper foretages, og ækvivalens eller noninferioritet kan kun påvises, hvis begge analyser understøtter det. For at sikre den bedst mulige kvalitet af analysen er det vigtigt at anvende fuldstændige followupdata på alle randomiserede patienter svarende til per

protokol, uanset om de efterfølgende viser sig ikke at have opfyldt inklusionskriterierne, udgik førtidigt eller brød protokollen på anden måde. En sådan omhyggelig dataindsamling tillader maksimal fleksibilitet ved den senere analyse og giver derfor et mere robust resultat.

Det mest almindelige problem i de publicerede ækvivalens- og noninferioritetsundersøgelser er, at de planlægges og analyseres, som om de var superioritetsundersøgelser, og at manglen på en statistisk signifikant forskel tages som bevis på ækvivalens. Nylige publikationer vedrørende kvaliteten af gennemførte RCT [6, 7] tyder på, at der er behov for et bedre kendskab til, hvordan ækvivalens- og noninferioritetsundersøgelser skal planlægges, udføres, analyseres og rapporteres.

Korrespondance: Erik Christensen, Intern Medicinsk Klinik I, Bispebjerg Hospital, DK-2400 København NV. E-mail: ec05@bbh.regionh.dk

Antaget: 29. april 2008  
Interessekonflikter: Ingen

### Litteratur

1. Kampmann JP, Christensen E. Den randomiserede kliniske undersøgelse. I: Jørgensen T, Christensen E, Kampmann JP, eds. Klinisk forskningsmetode. 2. udg. København: Munksgaard, 2005:19-32.
2. Christensen E. Methodology of superiority vs. equivalence trials and non-inferiority trials. *J Hepatol* 2007;46:947-54.
3. Christensen E, Kampmann JP. Statistik. I: Jørgensen T, Christensen E, Kampmann JP, eds. Klinisk forskningsmetode. 2. udg. København: Munksgaard, 2005:149-93.
4. Jones B, Jarvis P, Lewis JA et al. Trials to assess equivalence: the importance of rigorous methods. *BMJ* 1996;313:36-39.
5. Wiens BL. Choosing an equivalence limit for noninferiority or equivalence studies. *Control Clin Trials* 2002;23:2-14.
6. Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJW, CONSORT Group. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. *JAMA* 2006;295:1152-60.
7. Le Henanff A, Giraudeau B, Baron G, Ravaud P. Quality of reporting of noninferiority and equivalence randomized trials. *JAMA* 2006;295:1147-51.