9. Cox CE, Salud CJ, Cantor A et al. Learning curves for breast cancer sentinel lymph node mapping based on surgical volume analysis. J Am Coll Surg 2001;193:593-600.
10. Chagpar AB, Martin RC, Scoggins CR et al. Factors predicting failure to identify a sentinel lymph node in breast cancer. Surgery 2005;138:56-63.
11. Klauber-DeMore N, Tan LK, Liberma L et al. Sentinel Lymph Node Biopsy: Is it indicated in patients with high-risk ductal carcinoma-in-situ and ductal carcinoma-in-situ with microinvasion? Ann Surg Oncol 2000;7:636-42.
12. Yen TWF, Hunt KK, RossMI et al. Predictors of invasive breast cancer in patients with an initial diagnosis of ductal carcinoma in situ: a guide to selective management of ductal carcinoma in situ. J Am Coll Surg 2005; 200:516-26
13. Dupont EL, Kamath VJ, Ramnath EM et al. The role of lymphoscintigraphy in the management of the patient with breast cancer. Ann Surg Oncol 2001;8: 354-60.
14. Tuttle TM, Zogakis TG, Dunst CM et al. A review of technical aspects of sentinel lymph node identification for breast cancer. Am Coll Surg 2002;195:261-8.
15. Shimazu K, Tamaki Y, Tagichi T et al. Comparison between periareolar and peritumoral injection of radiotracer for sentinel lymph node biopsy in patients with breast cancer. Surg 2002;31:277-86.
16. Wong SL, Cao C, Edwards MJ et al. Accuracy of sentinel lymph node biopsy for patients with T2 and T3 breast cancers. Am Surg 2001;67:522-6.
17. Naik AM, Fey J, Gemignani M et al. The risk of axillary relapse after sentinel lymph node biopsy for breast cancer is comparable with that of axillary lymph node dissection. Ann Surg 2004;240:462-71.

Er relative risici og odds-ratioer i resumeer troværdige? Sekundærpublikation

Overlæge Peter C. Gøtzsche

H:S Rigshospitalet, Det Nordiske Cochrane Center

Resume

Det første resultat i 520 resumeer, hvori man angav relative risici eller odds-ratioer, var statistisk signifikant i 70% af de randomiserede forsøg, 84% af kohortestudierne og 84% af case-kontrol-studierne. Mange resultater kom fra subgruppeanalyser eller sekundære analyser, eller biasede udvalg af resultater. Fordelingen af p-værdier i intervallet 0,04-0,06 var ekstremt skæv for randomiserede forsøg, men de fleste signifikante resultater var forkerte, meget tvivlsomme eller kunne diskuteres. Skævheden var endnu større i observationelle studier. Signifikante resultater i resumeer bør tolkes meget varsomt.

I en forskningsartikel læser de fleste kun resumeet. Det er derfor vigtigt, at resumeet klart afspejler studiet og præsenterer resultaterne på en balanceret måde. Dette er ikke altid tilfældet. I en undersøgelse blev det fundet, at bias i konklusionen eller resumeet af sammenlignende forsøg med nonsteroide antiinflammatoriske præparater konsekvent favoriserede det nye præparat frem for kontrolpræparatet i 81 forsøg og favoriserede kontrolpræparatet i kun et forsøg [1]. I en anden undersøgelse af 73 observationelle studier blev der fundet en overvægt af p-værdier mellem 0,01 og 0,05 i resumeerne, hvilket tyder på bias i analysen eller rapporteringen [2].

Jeg undersøgte i et stort udvalg af forskningsartikler, om p-værdier i resumeer generelt er troværdige.

Metode

Fordelingen af p-værdier i resumeer af randomiserede forsøg og observationelle studier blev sammenlignet, og årsagerne til eventuelle skævheder, især for p-værdier, som lå tæt på det konventionelle signifikansniveau, $p < 0,05$, blev undersøgt.

Jeg søgte på PubMed efter resumeer af artikler, som var publiceret i 2003 og indeholdt ordene *relative risk* eller *odds ratio* i et hvilket som helst felt. Der var 260 randomiserede forsøg, hvori man rapporterede om mindst et binært effektmål, og jeg indsamlede derefter en stikprøve på 130 kohortestudier og 130 case-kontrol-studier ud fra en randomiseringsliste [3].

Den først rapporterede relative risiko eller odds-ratio og p-værdien blev benyttet. Hvis der ikke var nogen p-værdi, beregnede jeg den fra sikkerhedsintervallet, når dette forelå, ved at anvende normalfordelingstilnærmelsen efter logtransformation. Hvis det første resultat ikke var statistisk signifikant, noteredes, om der var nogen signifikante resultater i resten af resumeet.

Fordelingerne af p-værdier mellem forsøg og observationelle studier og mellem kohortestudier og case-kontrol-studier blev sammenlignet med Mann-Whitney-testen efter kategorisering [2].

Endelig undersøgte jeg, om rapporterede p-værdier i intervallet 0,04-0,06 var korrekte ved at sammenligne med metode- og resultatafsnittet efter at have skaffet hele artiklen. Alle data blev dobbelttjekket [3] for at reducere risikoen for taste- eller fortolkningsfejl.

Resultater

Det først rapporterede binære effektmål i resumeet var den relative risiko i 52% af de randomiserede forsøg, 35% af kohortestudierne og 4% af case-kontrol-studierne. Dette resultat var statistisk signifikant ($p < 0,05$) i 70% af forsøgene, i 84% af kohortestudierne og i 84% af case-kontrol-studierne. P-værdierne var mere ekstreme i observationelle studier end i randomiserede forsøg ($p < 0,001$) og mere ekstreme i kohorte-

studier end i case-kontrol-studier (p = 0,04). Når man inddrog samtlige resultater i resumeerne, blev der i hhv. 86%, 93% og 93% af studierne rapporteret om signifikante resultater.

Fordelingen af p-værdier i intervallet 0,04-0,06 var ekstremt skæv. Man ville forvente, at antallet af p-værdier i intervallet $0,05 \leq p < 0,06$ svarede til antallet i intervallet $0,04 \leq p < 0,05$, men jeg fandt hhv. fem og 46, hvilket er et højst usandsynligt fund ($p < 0,0001$), hvis man antager, at forskere ikke er biasede, når de analyserer og rapporterer om data.

Kun fem randomiserede forsøg havde $0,05 \leq p < 0,06$, hvorimod 29 forsøg havde $0,04 \leq p < 0,05$. Jeg kunne kontrollere beregningerne for hhv. fire og 23 af disse forsøg og bekræftede tre af de fire ikkesignifikante resultater. Det fjerde resultat var anført som p = 0,05, hvilket forfatterne tolkede som signifikant; jeg fik p = 0,03. Otte af de 23 signifikante resultater var korrekte, fire var forkerte, fem var meget tvivlsomme, fire kunne diskuteres, og to var kun signifikante, hvis en χ^2-test uden kontinuitetskorrektion blev benyttet [3].

Fordelingen af p-værdier i intervallet 0,04-0,06 var endnu mere ekstrem for observationelle studier. I ni kohortestudier og otte case-kontrol-studier blev der rapporteret om p-værdier i dette interval, men i samtlige 17 tilfælde var $p < 0,05$. Analyserne var justeret for konfoundere, og resultaterne kunne derfor ikke genregnes.

Diskussion

Det var uventet, at man i så mange resumeer af randomiserede forsøg præsenterede signifikante resultater, eftersom *clinical equipoise*, dvs. det forhold, at man ikke kan sige noget sikkert om, hvilken behandling der er bedst, inden forsøget går i gang, er en etisk forudsætning for at udføre randomiserede forsøg. Dertil kommer, at den statistiske styrke i mange forsøg er meget lille [4], hvorfor det er svært at forkaste nulhypotesen om ingen forskel.

Vor igangværende forskning har vist, at mere end 200 statistiske test undertiden planlægges i forsøgsprotokoller. Hvis man sammenligner en behandling med sig selv, dvs. nulhypotesen om ingen forskel er sand, er chancen 99,996% ($= 1-0,95^{200}$) for, at en eller flere ud af 200 test vil være statistisk signifikante på 5%-niveauet, hvis vi antager, at der er tale om uafhængige test. Forskeren eller sponsor kan således være temmelig sikker på, at »noget interessant vil vise sig«. Der tages sjældent behørigt hensyn til multiple signifikanstest, og det er i reglen ikke muligt at skelne troværdigt mellem primære og sekundære effektmål. I et studie, hvori vi sammenlignede protokoller med forsøgsrapporter, påviste vi selektiv publicering af resultater, afhængigt af de opnåede p-værdier, og at mindst et primært effektmål var ændret, indført eller udeladt i 62% af forsøgene [5].

Der er også et stort spillerum for bias i observationelle studier. Mange studier er for små, og der rapporteres ikke om statistiske styrkeberegninger [2]. Desuden viste en undersøgelse, at man i 92% af artiklerne korrigerede for konfoundere med en median på syv, men man angav i reglen ikke, om valg af konfoundere var besluttet på forhånd [2]. I 14% af disse artikler rapporterede man over 100 effektestimater, og i 57% af forsøgene var der subgruppeanalyser, hvis resultater forskerne generelt troede på [2].

Den manglende randomisering i observationelle studier betyder, at en næsten hvilken som helst sammenligning bliver statistisk signifikant, hvis materialet er stort nok, fordi de sammenlignede grupper næsten altid vil være forskellige [6]. Derfor er p-værdier særligt misvisende i observationelle studier og burde slet ikke tolkes som sandsynligheder [6]. Dette grundlæggende problem er formentlig en af årsagerne til, at p-værdier i kohortestudier er mest ekstreme, idet data fra mange store kohortestudier publiceres meget ofte [2].

Da påståede årsags-virknings-sammenhænge så ofte er falsk alarm, har nogle erfarne epidemiologer den tommelfingerregel ikke at lade sig anfægte af skadevirkninger påvist i observationelle studier, medmindre risikoen er forøget tre gange [7]. Dette tal skal helst ligge uden for sikkerhedsintervallet, idet selv en odds-ratio på 20,5 falmer, hvis det viser sig, at sikkerhedsintervallet går fra 2,2 til 114,0. Sikkerhedsintervaller forelå eller kunne beregnes for det første resultat i 116 resumeer af case-kontrol-studierne, men kun i seks tilfælde (5%) var der en vis sikkerhed for en tre gange forøget risiko.

Selv om der var mange signifikante resultater i resumeerne, var disse stærkt selektive, f.eks. *»The strongest mechanical risk factor«, »The only factor associated with ...«, »The highest odds ratio ...«*, og kun i få resumeer tog man forbehold over for disse data. Jeg gennemgik de 181 signifikante resumeer af randomiserede forsøg endnu engang, men fandt kun fire forbehold (2%), selv om subgruppeanalyser eller sekundære analyser og korrigering for konfoundere i regressionsanalyser var almindelige, hvilket også den hyppige brug af odds-ratio frem for relativ risiko tydede på [3]. I overensstemmelse hermed blev det i en anden oversigt påvist, at de fleste resultater af subgruppeanalyser i randomiserede forsøg fandt vej til resumeet eller konklusionen i artiklerne [8].

Jeg ville undersøge bias under dataanalysen nøjere og fokuserede på p-værdier mellem 0,04 og 0,06, selv om p-værdier i dette interval ud fra et statistisk synspunkt selvfølgelig bør tolkes på samme måde. Nogle af de signifikante resultater var forkerte eller meget tvivlsomme. Dette stemmer overens med en undersøgelse af lægemiddelforsøg, hvori det i reglen ikke var muligt at kontrollere beregningerne [1]. Imidlertid fandt jeg ti forsøg, hvori de signifikante resultater var forkerte, og jeg havde en stærk mistanke om falsk positive resultater i yderligere fem forsøg; i samtlige tilfælde blev det nye præparat favoriseret frem for kontrolpræparatet [1].

Min konklusion er, at man generelt ikke kan stole på signifikante resultater i resumeer. Hvis der overhovedet er behov for et konventionelt signifikansniveau, hvilket er tvivlsomt, kunne overvægten af signifikante resultater reduceres, hvis: 1) signifikansniveauet var $p < 0,001$, som det er foreslået for

observationelle studier [9], 2) dataanalysen og udarbejdelse af manuskriptet blev udført blindt, uden kendskab til de afgørende forhold, f.eks. typen af intervention, eksponering eller sygdomsstatus, før alle forfattere havde godkendt de to versioner af manuskriptet [10], og 3) tidsskriftsredaktørerne foretog en mere kritisk gennemgang af resumeerne og krævede, at forskningsprotokoller og rådata – både for randomiserede forsøg og for observationelle studier – blev indsendt sammen med manuskriptet.

Korrespondance: *Peter C. Gøtzsche*, Det Nordiske Cochrane Center, Afdeling 7112, H:S Rigshospitalet, DK-2100 København Ø. E-mail: pcg@cochrane.dk

Antaget: 24. februar 2006
Interessekonflikter: Ingen angivet

This article is based on a study first reported in the BMJ 2006;333:231-4.

Litteratur

1. Gøtzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal, antiinflammatory drugs in rheumatoid arthritis. Controlled Clin Trials 1989;10:31-56, erratum:1989;10:356.
2. Pocock SJ, Collier TJ, Dandreo KJ et al. Issues in the reporting of epidemiological studies: a survey of recent practice. BMJ 2004;329:883.
3. Gøtzsche PC. Believability of relative risks and odds ratios in abstracts: cross sectional study. BMJ 2006;333:231-4.
4. Chan AW, Altman DG. Epidemiology and reporting of randomised trials published in PubMed journals. Lancet 2005;365:1159-62.
5. Chan A-W, Hróbjartsson A, Haahr MT et al. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 2004;291:2457-65.
6. Greenland S. Randomization, statistics, and causal inference. Epidemiology 1990;1:421-9.
7. Taubes G. Epidemiology faces its limits. Science 1995;269:164-9.
8. Assmann SF, Pocock SJ, Enos LE et al. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000;355:1064-9.
9. Smith GD, Ebrahim S. Data dredging, bias, or confounding. BMJ 2002;325:1437-8.
10. Gøtzsche PC. Blinding during data analysis and writing of manuscripts. Controlled Clin Trials 1996;17:285-90.

Myokardienekrose med sen forekomst af ventrikulær takyarytmi sekundært til diabetisk ketoacidose

Reservelæge Søren Kildeberg Paulsen,
1. reservelæge Steen Hvitfeldt Poulsen, overlæge Leif Thuesen & overlæge Jens Friis Bak

Århus Sygehus, Medicinsk-endokrinologisk Afdeling C, og Skejby Sygehus, Kardiologisk Afdeling B

Diabetisk ketoacidose er en alvorlig komplikation hos insulinbehandlede diabetikere og skyldes enten absolut eller relativ insulinmangel, sidstnævnte ofte grundet øget insulinresistens ved infektion, akut myokardieinfarkt mv. Diabetisk ketoacidose er en potentielt livstruende tilstand. I nærværende sygehistorie omtales en ung kvinde med svær diabetisk ketoacidose, der trods korrektion af elektrolytter og syre-base-status, får tegn på ST-elevationsmyokardieinfarkt og ventrikulær takykardi 27 timer efter indlæggelsen.

Sygehistorie

En 27-årig kvinde, der havde haft type 1-diabetes gennem otte år, blev indlagt akut, efter at hun blev fundet ukontaktbar i sit hjem. I dagene op til indlæggelsen havde hun været plaget af kvalme, opkastninger og diare. Patienten havde undladt at tage insulin som følge af manglende evne til at spise. Ved indlæggelsen var hun forpint og bevidsthedspåvirket. Temperaturen var 36,1 °C, blodtrykket 130/75 mmHg og pulsen 100/min. Arteriel blodgasanalyse viste svær metabolisk acidose med pH 6,90; pCO_2 1,2 kPa; pO_2 16,9 kPa, umålelig lavt standardbikarbonat og baseoverskud på -27,9. Plasma (P)-kreatinin var 228 µmol/l; P-kalium 7,3 mmol/l; P-natrium: 128 mmol/l; P-glukose: 63,6 mmol/l; 3-OH-butyrat på 5,2 mmol/l; leukocytter 28,6 mia./l samt C-reaktivt protein (CRP) 450 nmol/l. Ketoacidosen rettede sig under behandling på intensiv afdeling med kontinuerlig intravenøs infusion af hurtigtvirkende insulin 6 IE pr. time, isoton natriumchlorid, kaliumchlorid og glukose i variabel infusionshastighed efter afdelingens ketoacidoseregimen. P-kalium blev normaliseret inden for to timer og forblev i normalområdet under resten af forløbet. Som følge af vedvarende bevidsthedspåvirkning 12 timer efter indlæggelsestidspunktet blev der foretaget computertomografi af cerebrum og lumbalpunktur. Begge var uden patologiske fund, fraset let cerebralt ødem. Omkring 27 timer efter indlæggelsestidspunktet opstod der ST-segment-forandringer og flere kortvarige løb af ventrikulær takykardi på overvågningstelemetri (**Figur 1**). Tolv-aflednings-elektrokardiogram (EKG) viste ST-elevationer i II, III, AvF og i alle prækordialafledninger. Biokemisk var elektrolytterne da normaliserede, og acidosen var svundet med normalt pH og HCO_3^-. Patienten havde ikke haft brystsmerter eller palpitationsfornemmelser. Efter yderligere 12 timer bemærkedes aftagende