BMJ

BMJ 2011;343:d3340 doi: 10.1136/bmj.d3340

RESEARCH METHODS & REPORTING

Interpreting and reporting clinical trials with results of borderline significance

Borderline significance in the primary end point of trials does not necessarily mean that the intervention is not effective. Researchers and journals need to be more consistent in how they report these results

Allan Hackshaw *deputy director*, Amy Kirkwood *statistician*

Cancer Research UK and UCL Cancer Trials Centre, University College London, London W1T 4TJ

The quality of randomised clinical trials and how they are reported have improved over time, with clearer guidelines on conduct and statistical analysis.[1] Clinical trials often take several years, but interpreting the results at the end is arguably the most important activity because it influences whether a new intervention is recommended or not. Although researchers have become more familiar with medical statistics, the interpretation and reporting of results of borderline significance remains a problem. We examine the problem and recommend some solutions.

What is the problem?

New interventions used to be compared with minimal or no treatment, so researchers were looking for and finding large treatment effects. Clear recommendations were made because the P values were usually small (eg, P<0.001). However, modern interventions are usually compared with the existing standard treatment, so that the effects are often expected to be smaller than before, and it is no longer as easy to get small P values. The cut-off used to indicate a real effect is widely taken as P=0.05 (called statistically significant). The problem is that although P=0.05 is an arbitrary figure, many researchers still adhere strictly to it when making conclusions about an intervention, and often use it as the sole basis for this. Researchers and journals sometimes conclude that there is no effect.

The P=0.05 cut-off was first proposed by R A Fisher in 1925 as being low enough to make decisions, and over time has become widely adopted. However, examining interventions with P values just above 0.05 is difficult, especially if the trial is unique. It is incorrect to regard, for example, a relative risk of 0.75 with a 95% confidence interval of 0.57 to 0.99 and P=0.048 as clear evidence of an effect, but the same point estimate with a 95% confidence interval of 0.55 to 1.03 and P=0.07 as showing no effect, simply because one P value is just below 0.05 and the other just above. Although the issue has been raised before,[2 3] it still occurs in practice.

P values are an error rate (like the false positive rate in medical screening). In the same way that a small P value does not guarantee that there is a real effect, a P value just above 0.05 does not mean no effect. If P=0.049, we expect to claim that a new intervention is beneficial, when it really is not, almost 5% of the time, but the intervention would probably still be recommended. The size of a P value depends on two factors: the magnitude of the treatment effect (relative risk, hazard ratio, mean difference, etc) and the size of the standard error (which is influenced by the study size, and either the number of events or standard deviation, depending on the type of outcome measure used). Very small P values (the easiest to interpret) arise when the effect size is large and the standard error is small. Borderline P values can occur when there is a clinically meaningful treatment effect but a large or moderate standard error—often because of an insufficient number of participants or events (the trial is referred to as being underpowered). This is perhaps the most common cause of borderline results. Borderline P values can also occur when the treatment effect is smaller than expected, which with hindsight would have a required a larger trial to produce a P value <0.05, so again the study is underpowered.

Using confidence intervals

Confidence intervals are usually more informative than the P value when borderline results are found, as the following example shows. The EICESS-92 phase III trial aimed to determine whether adding etoposide to standard ifosfamide chemotherapy would improve event-free survival in Ewing's sarcoma.[4] Powered to detect a hazard ratio of 0.60 (40% relative risk reduction), the target sample size was 400 patients (492 were recruited). The observed hazard ratio was 0.83 (95%

Correspondence to: Allan Hackshaw ah@ctc.ucl.ac.uk

Extra material supplied by the author (see http://www.bmj.com/content/343/bmj.d3340/suppl/DC1)

confidence interval 0.65 to 1.05, P=0.12), Because P>0.05 it would normally be concluded that there is insufficient evidence for an effect, even though the 17% risk reduction is clinically important, but smaller than the 40% expected. Most researchers and journal reviewers understand that the true effect is likely to lie somewhere in the confidence interval range, hence the possibility of it being 1.0—that is, no effect. However, there is a common misconception that the true effect lies anywhere within this range with equal likelihood. It is more likely to be around the estimated hazard ratio (0.83, the best estimate of the true effect) than at either extremes of the confidence interval. Thus, although the upper limit (1.05) is just above the no effect value, there is only a 6% chance that it exceeds 1.0 (figure). There is a 50% chance that the true hazard ratio is between 0.77 and 0.90, or 75% chance that it is between 0.72 and 0.95; therefore a treatment benefit is likely. The authors concluded that "the addition of etoposide seemed to be beneficial." This is appropriate wording because it is the only randomised study to evaluate adding etoposide to an ifosfamide regimen in this patient group, and the disorder is uncommon (it took 6.5 years to recruit 492 patients). Even 7.5 years after recruitment had ended the number of events (n=266) still did not allow the primary end point to have P<0.05. To conclude insufficient evidence or, worse still, no effect, would have been incorrect and a useful result from a unique trial would be missed. Although the target sample size was exceeded, the observed treatment effect was smaller than originally expected, hence the lack of statistical significance.

Inconsistency in language in clinical trial reports

We examined the *BMJ*, *Lancet*, *JAMA*, *New England Journal of Medicine*, *Journal of the National Cancer Institute*, and the *Journal of Clinical Oncology* to see how the results or conclusions of randomised phase III trials published in 2009 were described in the abstract (which most readers focus on). Out of 287 studies, 24 (1 in 12) were considered to have borderline results when the direction of the primary end point indicated a treatment benefit, with P value between 0.05 and 0.10 or a lower or upper 95% confidence limit close to the no effect value (that is, 1 for risk or hazard ratios and 0 for risk or mean differences). The table gives examples and a full list is available on bmj.com.[5-28] There is a general inconsistency in the language used. The intention here is not to alter the published conclusions, but only to be aware of the differences in how they were reported.

Among 10 article abstracts that concluded or gave the impression of no effect for the primary end point, seven had P values of 0.11 to 0.17 with hazard or odds ratios ranging from 0.85 to 0.90 and upper 95% confidence limits of 1.02 to 1.07, and one had a mean difference of 0.06×10^3 (95% confidence interval -0.002 to 0.13×10^3) with a P value of 0.06.[5] These results suggest that there is probably some effect, but perhaps not clinically worthwhile. However, a seemingly large effect was found in two trials with P=0.06 and upper 95% confidence limits just above the no effect value (table). One had a mean difference in scores of −27.8 units[7] and the other a relative risk of 0.66,[11] so there probably is a real benefit on the primary end point in both cases.

Eleven articles concluded that there was a suggestion of an effect, usually with moderate to large treatment effects, and P values 0.06 to 0.10 (often 0.06 or 0.07). However, in one trial the effect seemed relatively small (hazard ratio=0.93, 95% confidence interval 0.84 to 1.02, P=0.13[19]). All of these articles seemed to base their conclusions not just on the P value but on other end points or adjusted results. In two trials, the achieved sample size was much lower than the target because of poor accrual, but both found large treatment effects (hazard ratios 0.67 and 0.69), which would probably have been significant if there had been more patients.[22 24]

Three articles concluded an effect with some confidence. The treatment effects were variable, and P values ranged between 0.06 and 0.1. Again, authors sometimes drew attention to other significant end points but the language in relation to the main outcome measure could have been less strong in some cases.

Although seven of the 24 studies were large (≥2000 participants), this did not guarantee clear results for the primary end point. The overall conclusions of several studies were often supported by results for other end points or from other trials, and the possibility of an effect was discussed outside the abstract. However, it is inconsistent that, for example, two trials with similar effect sizes (risk ratios of 0.66 and 0.67) and P values (0.06 and 0.07) came to different conclusions (no effect [11] and suggestion of an effect[24]), and two trials with a smaller effect size (hazard ratio 0.84) but similar or larger P value (0.06 or 0.13) indicated a possible effect (table).[19 25] It is also useful to consider borderline confidence intervals and P values when a trial intervention unexpectedly suggests harm in relation to the primary end point. Authors might be more inclined to make firmer conclusions in this situation than if an intervention shows evidence of benefit. However, we found two examples where there were more events in the intervention group but the authors concluded only that there was no benefit. In a randomised trial of 635 patients with type 2 diabetes and diabetic retinopathy, the primary end point was developing clinically important macular oedema[29]; the hazard ratio for calcium dobesilate versus placebo was 1.32 (0.96 to 1.81, P=0.08), but the conclusion in the abstract stated only that calcium dobesilate did not reduce the risk of developing macular oedema. Similarly, in a trial of 486 head and neck cancer patients comparing gefitinib (250 or 500 mg) with methotrexate,[30] the hazard ratios for mortality were 1.22 (P=0.12) and 1.12 (P=0.39) for 250 and 500 mg, respectively. The conclusion reported in the abstract stated that "neither gefitinib 250 nor 500 mg/day improved overall survival," though the pooled hazard ratio would be 1.17 (95% confidence interval 0.98 to 1.39).

Possible solutions

The problem of borderline results could be avoided by designing trials with small or moderate effect sizes. However, this is often not feasible because large sample sizes are usually required, which is particularly challenging in uncommon disorders. But even with careful trial design and good prior evidence, the observed treatment effect can be noticeably lower than that expected, thus producing P values that are just above 0.05 (as in the Ewing's sarcoma trial above). A possible solution is to use a validated and established surrogate marker as the primary (or co-primary) end point—for example, progression-free survival instead of overall survival in some cancer trials, or cholesterol for some prevention trials in cardiovascular disease. There should be more events if a surrogate marker is used, and this will increase the chance of the result being statistically significant. However, researchers need to be aware that finding significant results for a true end point (such as survival) will be difficult because such studies are smaller. Furthermore, borderline results could still be found with any end point.

Meta-analysis can also be a solution, but only if there are two or more trials to combine. This was indeed the case for one of

the articles we found,[15] where the hazard ratio for one trial was 0.86, 95% confidence interval 0.72 to 1.02 (P=0.08), but the pooled effect from three trials was 0.86, 0.75 to 0.98 (P=0.02). However, there are many instances when the trial is the only one and will not be repeated, usually because of greater interest in newer interventions or because limitations of sample size or rarity of the disorder make it unfeasible to repeat the trial. Unique trials might become more common because international clinical trial registers now allow researchers to check if similar studies are in progress elsewhere. Although it is generally good practice to have at least two trials of the same intervention (with consistent results) before recommending it for routine use, researchers might be less inclined to conduct a replicate trial or be less likely to receive a grant from funding organisations.

Recommendations

The figure shows why a confidence interval is a better way of interpreting data when borderline results are found. Importantly, it shows that even when P>0.05, there is a higher likelihood that the true effect lies around the point estimate from the trial, rather than at the ends of the confidence interval, so a treatment effect should not be readily dismissed if it seems clinically meaningful.

Borderline results cannot be used as strong evidence either for or against an intervention. If a clinically important effect is observed with a P value just above 0.05 (or an upper or lower confidence limit close to the no effect value), it is incorrect to conclude no effect and not consider further what is likely to be an effective intervention, especially for uncommon disorders or trials that took many years to complete. Researchers should examine other end points to look for consistency and other evidence (for example, cohort studies, dose-response relations, or similar types of treatments that show a clear effect). Importantly, they should state that there is evidence for the primary end point but use moderate words such as "suggestion," "seems," or "indication" that need to be accepted consistently by journals. If the treatment effect is lower than expected, the clinical implications could be specifically discussed. The aim is to avoid giving the reader the impression that an intervention is completely ineffective, when it is likely to be effective. Similarly, researchers of trials with results of borderline significance should not give the impression that the evidence is conclusive. The same principles apply to other areas of research such as examining risk factors for and causes of disorders or early death.

We thank Nicholas Wald for his helpful comments.

Competing interest: All authors have completed the ICJME unified disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare no support from any organisation for the submitted work; no financial relationships with any organisation that might have an interest in the submitted work in the previous three years; and no other relationships or activities that could appear to have influenced the submitted work.

Contributors: AH had the original idea, AK reviewed the journals to identify clinical trial reports that had borderline results, and both authors reviewed the articles, wrote the paper, and approved the final version. AH is the guarantor.

1 Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Ann Intern Med* 2001;134:657-62.
2 Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ* 1995;311:485.
3 Alderson P. Absence of evidence is not evidence of absence. *BMJ* 2004;328:476.
4 Paulussen M, Craft AW, Lewis I, Hackshaw A, Weston C, Douglas C, et al. The European Intergroup Cooperative Ewing's Sarcoma Study Group. EICESS-92—results of two randomised trials of the European Intergroup Cooperative Ewing's Sarcoma Study. *J Clin Oncol* 2008;26:4385-93.
5 ASTRAL Investigators. Revascularization versus medical therapy for renal-artery stenosis. *N Engl J Med* 2009;361:1953-61.
6 Azzopardi DV, Strohm B, Edwards AD, Dyet L, Halliday HL, Juszczak E, et al. Moderate hypothermia to treat perinatal asphyxial encephalopathy. *N Engl J Med* 2009;361:1349-58.
7 Bakitas M, Lyons KD, Hegel MT, Balan S, Brokaw FC, Seville J, et al. Effects of a palliative care intervention on clinical outcomes in patients with advanced cancer: the Project ENABLE II randomized controlled trial. *JAMA* 2009;302:741-9.
8 Bhatt DL, Lincoff AM, Gibson CM, Stone GW, McNulty S, Montalescot G, et al. Intravenous platelet blockade with cangrelor during PCI. *N Engl J Med* 2009;361:2330-41.
9 Duckworth W, Abraira C, Moritz T, Reda D, Emanuele N, Reaven PD, et al. Glucose control and vascular complications in veterans with type 2 diabetes. *N Engl J Med* 2009;360:129-39.
10 Mehta SR, Granger CB, Boden WE, Steg PG, Bassand JP, Faxon DP, et al. Early versus delayed invasive intervention in acute coronary syndromes. *N Engl J Med* 2009;360:2165-75.
11 Murphy AW, Cupples ME, Smith SM, Byrne M, Byrne MC, Newell J, et al. Effect of tailored practice and patient care plans on secondary prevention of heart disease in general practice: cluster randomised controlled trial. *BMJ* 2009;339:b4220.
12 Lefebvre JL, Rolland F, Tesselaar M, Bardet E, Leemans CR, Geoffrois L, et al. Phase 3 randomized trial on larynx preservation comparing sequential vs alternating chemotherapy and radiotherapy. *J Natl Cancer Inst* 2009;101:142-52.
13 Löwenberg B, Ossenkoppele GJ, van Putten W, Schouten HC, Graux C, Ferrant A, et al. High-dose daunorubicin in older patients with acute myeloid leukemia. *N Engl J Med* 2009;361:1235-48.
14 Van Cutsem E, Labianca R, Bodoky G, Barone C, Aranda E, Nordlinger B, et al. Randomised phase III trial comparing biweekly infusional fluorouracil/leucovorin alone or with irinotecan in the adjuvant treatment of stage III colon cancer: PETACC-3. *J Clin Oncol* 2009;27:3117-25.
15 Cunningham D, Chau I, Stocken DD, Valle JW, Smith D, Steward W, et al. Phase III randomized comparison of gemcitabine versus gemcitabine plus capecitabine in patients with advanced pancreatic cancer. *J Clin Oncol* 2009;27:5513-8.
16 Fidias PM, Dakhil SR, Lyss AP, Loesch DM, Waterhouse DM, Bromund JL, et al. Phase III study of immediate compared with delayed docetaxel after front-line therapy with gemcitabine plus carboplatin in advanced non-small-cell lung cancer. *J Clin Oncol* 2009;27:591-8.
17 Grau MV, Sandler RS, McKeown-Eyssen G, Bresalier RS, Haile RW, Barry EL, et al. Nonsteroidal anti-inflammatory drug use after 3 years of aspirin use and colorectal adenoma risk: observational follow-up of a randomized study. *J Natl Cancer Inst* 2009;101:267-76.
18 Hess G, Herbrecht R, Romaguera J, Verhoef G, Crump M, Gisselbrecht C, et al. Phase III study to evaluate temsirolimus compared with investigator's choice therapy for the treatment of relapsed or refractory mantle cell lymphoma. *J Clin Oncol* 2009;27:3822-9.
19 O'Connor CM, Whellan DJ, Lee KL, Keteyian SJ, Cooper LS, Ellis SJ, et al. Efficacy and safety of exercise training in patients with chronic heart failure: HF-ACTION randomized controlled trial. *JAMA* 2009;301:1439-50.
20 Plint AC, Johnson DW, Patel H, Wiebe N, Correll R, Brant R, et al. Epinephrine and dexamethasone in children with bronchiolitis. *N Engl J Med* 2009;360:2079-89.
21 Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S, Kaewkungwal J, Chiu J, Paris R, et al. Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. *N Engl J Med* 2009;361:2209-20.
22 Roh MS, Colangelo LH, O'Connell MJ, Yothers G, Deutsch M, Allegra CJ, et al. Preoperative multimodality therapy improves disease-free survival in patients with carcinoma of the rectum: NSABP R-03. *J Clin Oncol* 2009;27:5124-30.
23 Soligard T, Myklebust G, Steffen K, Holme I, Silvers H, Bizzini M, et al. Comprehensive warm-up programme to prevent injuries in young female footballers: cluster randomised controlled trial. *BMJ* 2008;337:a2469.
24 Stahl M, Walz MK, Stuschke M, Lehmann N, Meyer HJ, Riera-Knorrenschild J, et al. Phase III comparison of preoperative chemotherapy compared with chemoradiotherapy in patients with locally advanced adenocarcinoma of the esophagogastric junction. *J Clin Oncol* 2009;27:851-6.
25 Albain KS, Barlow WE, Ravdin PM, Farrar WB, Burton GV, Ketchel SJ, et al. Adjuvant chemotherapy and timing of tamoxifen in postmenopausal patients with endocrine-responsive, node-positive breast cancer: a phase 3, open-label, randomised controlled trial. *Lancet* 2009;374:2055-63.
26 Gomes MF, Faiz MA, Gyapong JO, Warsame M, Agbenyega T, Babiker A, et al. Pre-referral rectal artesunate to prevent death and disability in severe malaria: a placebo-controlled trial. *Lancet* 2009;373:557-66.
27 Jabre P, Combes X, Lapostolle F, Dhaouadi M, Ricard-Hibon A, Vivien B, et al. Etomidate versus ketamine for rapid sequence intubation in acutely ill patients: a multicentre randomised controlled trial. *Lancet* 2009;374:293-300.
28 Peterson AV Jr, Kealey KA, Mann SL, Marek PM, Ludman EJ, Liu J, et al. Group-randomized trial of a proactive, personalized telephone counseling intervention for adolescent smoking cessation. *J Natl Cancer Inst* 2009;101:1378-92.
29 Haritoglou C, Gerss J, Saverland C, Kampik A, Ulbig MW, et al. Effect of calcium dobesilate on occurrence of diabetic macular oedema (CALDIRET study): randomised, double-blind, placebo-controlled, multicentre trial. *Lancet* 2009;373:1364-71.
30 Stewart JSW, Cohen EEW, Licitra L, Van Herpen CML, Khorprasert C, Soulieres D, et al. Phase III study of gefitinib compared with intravenous methotrexate for recurrent squamous cell carcinoma of the head and neck. *J Clin Oncol* 2009;27:1864-71.

Accepted: 11 February 2011

Cite this as: *BMJ* 2011;343:d3340

Summary points

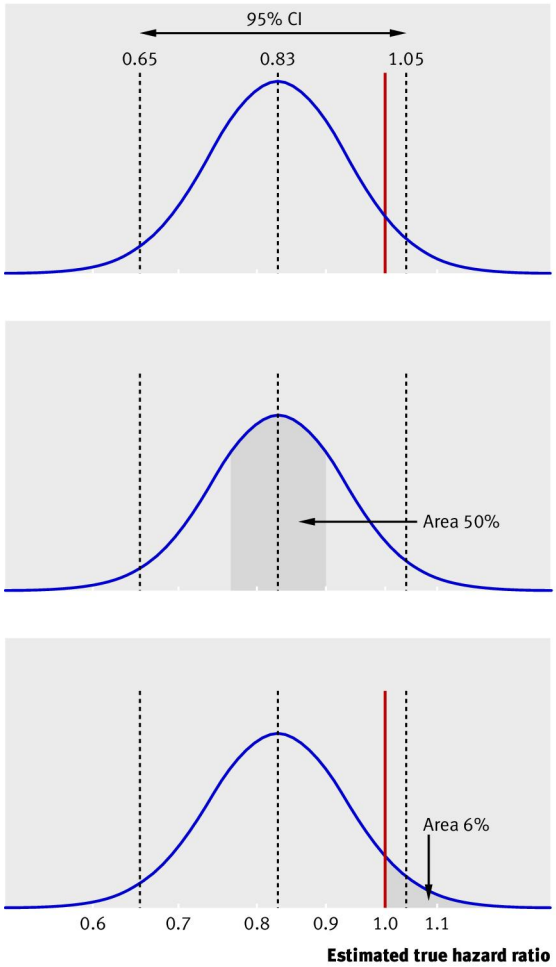
- Many researchers still adhere too strictly to the arbitrary P cut-off value of 0.05 when interpreting clinical trial results
- A P value that is just above 0.05 does not mean that there is no effect
- Confidence intervals are a better indicator of the likelihood of an effect and its size
- The true effect of an intervention is more likely to lie around the middle of a confidence interval (that is, the point estimate) than at either end
- Authors and journals should be more consistent in how they report primary trial end points of borderline significance

Table

Table 1| Examples of randomised phase III trials with borderline positive results but different author interpretations

Trial and author interpretation	Interventions and patient group (No of participants)	Primary end point	Main result (95% CI), P value	Conclusion in abstract
No effect on primary end point was concluded or implied				
Bakitas et al[7]	Nurse led psycho-educational intervention *v* usual care for palliative care in patients with advanced cancer (n=322)	Symptom intensity (quality of life and resource use were other end points)	Mean difference in score −27.8 (−57.2 to 1.6), P=0.06	"Those receiving nurse-led . . . intervention . . . had higher scores for quality of life and mood but did not have improvements in symptom intensity scores"
Murphy et al[11]	Tailored care plan *v* usual care in patients with coronary heart disease (n=903)	Patients with systolic blood pressure >140 mm Hg at 18 months (hospital admission was another end point)	Odds ratio 0.66 (0.43 to 1.01), P=0.06	"Admissions to hospital were significantly reduced…but no other clinical benefits were shown"
Lefebvre et al[12]	Alternating *v* sequential chemotherapy and radiotherapy in patients with resectable advanced squamous cell carcinoma of the larynx or hypopharynx (n=450)	Survival with a functional larynx	Hazard ratio 0.85 (0.68 to 1.06), P=0.15	"Larynx preservation, progression-free interval, and overall survival were similar in both arms"
Suggestion of effect on primary end point				
O'Connor et al[19]	Aerobic exercise training plus usual care *v* usual care alone, in patients with chronic heart failure (n=2331)	All cause mortality or hospital admission	Hazard ratio 0.93 (0.84 to 1.02), P=0.13	"Exercise training resulted in nonsignificant reductions in the primary endpoint"
Stahl et al[24]	Pre-surgical chemoradiotherapy *v* chemotherapy in patients with locally advanced cancer of the oesophagogastric junction (n=126, target was 576)	Overall survival	Hazard ratio 0.67 (0.41 to 1.07), P=0.07	"Although…statistical significance was not achieved, results point to a survival advantage for preoperative chemoradiotherapy"
Albain et al[25]	Cyclophosphamide, doxorubicin, and fluorouracil (CAF) followed by tamoxifen *v* CAF given concurrently with tamoxifen, in breast cancer (n=1116)	Disease-free survival	Hazard ratio 0.84 (0.70 to 1.01), P=0.06	"The adjusted HRs favoured CAF-T [sequential] over CAFT [concurrent] but did not reach significance for disease-free survival"
Clear effect on primary end point				
Jabre et al[27]	Ketamine versus etomidate among patients requiring sedation for emergency intubation (n=469)	Organ system function, determined from the maximum sequential organ failure assessment	Mean difference −0.7 (−1.4 to 0.0), P=0.056	"Our results show that ketamine is a safe and valuable alternative to etomidate"
Peterson et al[28]	Personalised telephone counselling using cognitive behavioural skills *v* no intervention to encourage smoking cessation in adolescents (n=2151)	6 months abstinence from smoking	Absolute risk difference 4.0% (−0.2% to 8.1%), P=0.06	"Personalized motivational interviewing … is effective in increasing teen smoking cessation"

Figure



Interpreting the results of a trial with hazard ratio 0.83 and 95% confidence interval 0.65 to 1.05. The vertical broken lines indicate the point estimate and the lower and upper confidence limits; the solid line is the no effect value