Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors

Asbjørn Hróbjartsson senior researcher1, Ann Sofia Skou Thomsen research associate1, Frida Emanuelsson research associate1, Britta Tendal postdoctoral fellow1, Jørgen Hilden associate professor of biostatistics2, Isabelle Boutron associate professor of epidemiology3, Philippe Ravaud professor of epidemiology3, Stig Brorson orthopaedic surgeon4

1Nordic Cochrane Centre, Rigshospitalet Department 3343, Blegdamsvej 9, 2100 Copenhagen Ø, Denmark; 2Department of Biostatistics, University of Copenhagen, Copenhagen; 3French Cochrane Centre, Assistance Publique (Hotel Dieu), INSERM U738, Université Paris Descartes, France; 4Department of Orthopaedic Surgery, Herlev University Hospital, Copenhagen

Abstract

Objective To evaluate the impact of non-blinded outcome assessment on estimated treatment effects in randomised clinical trials with binary outcomes.

Design Systematic review of trials with both blinded and non-blinded assessment of the same binary outcome. For each trial we calculated the ratio of the odds ratios—the odds ratio from non-blinded assessments relative to the corresponding odds ratio from blinded assessments. A ratio of odds ratios <1 indicated that non-blinded assessors generated more optimistic effect estimates than blinded assessors. We pooled the individual ratios of odds ratios with inverse variance random effects meta-analysis and explored reasons for variation in ratios of odds ratios with meta-regression. We also analysed rates of agreement between blinded and non-blinded assessors and calculated the number of patients needed to be reclassified to neutralise any bias.

Data Sources PubMed, Embase, PsycINFO, CINAHL, Cochrane Central Register of Controlled Trials, HighWire Press, and Google Scholar.

Eligibility criteria for selecting studies Randomised clinical trials with blinded and non-blinded assessment of the same binary outcome.

Results We included 21 trials in the main analysis (with 4391 patients); eight trials provided individual patient data. Outcomes in most trials were subjective—for example, qualitative assessment of the patient’s function. The ratio of the odds ratios ranged from 0.02 to 14.4. The pooled ratio of odds ratios was 0.64 (95% confidence interval 0.43 to 0.96), indicating an average exaggeration of the non-blinded odds ratio by 36%. We found no significant association between low ratios of odds ratios and scores for outcome subjectivity (P=0.27); non-blinded assessor’s overall involvement in the trial (P=0.60); or outcome vulnerability to non-blinded patients (P=0.52). Blinded and non-blinded assessors agreed in a median of 78% of assessments (interquartile range 64-90%) in the 12 trials with available data. The exaggeration of treatment effects associated with non-blinded assessors was induced by the misclassification of a median of 3% of the assessed patients per trial (1-7%).

Conclusions On average, non-blinded assessors of subjective binary outcomes generated substantially biased effect estimates in randomised clinical trials, exaggerating odds ratios by 36%. This bias was compatible with a high rate of agreement between blinded and non-blinded outcome assessors and driven by the misclassification of few patients.

Introduction

The randomised clinical trial is regarded as the most valid method for assessing the benefits and harms of healthcare interventions.1 One challenge to the validity of such trials is the tendency for assessments of outcomes to systematically deviate from the truth because of predispositions in observers, such as from hope or expectation.2Such observer bias, also called ascertainment bias or detection bias, might be especially important when outcome assessors have strong predispositions and when outcomes are subjective—that is, involve personal judgment such as with qualitative scores or pattern recognition of images. Similarly, observer bias might have little practical importance when neutral assessors evaluate an objective outcome, such as death.
Many trials use blinded outcome assessors to avoid bias, though use of non-blinded outcome assessors is also common, especially in non-pharmacological trials. For example, one study of orthopaedic trauma trials reported that blinded outcome assessment had not been implemented in 90% of trials. It is an empirical question to which degree the estimated effects of experimental interventions in randomised trials are affected by lack of blinding of the outcome assessors and which factors influence the degree of bias.

The most reliable way of studying the impact of non-blinded outcome assessors is to analyse trials that use both blinded and non-blinded assessors for the same outcome. One such trial by Noseworthy and colleagues is often cited, reporting that the effect of plasma exchange for multiple sclerosis was significant only with assessments by non-blinded neurologists. The finding, however, was inconsistent across time points, seen only for one of the two experimental interventions, and might be atypical. Other studies have been based on indirect comparisons with a considerable risk of confounding.

It is prudent to suspect possible bias in trials with non-blinded assessors. Existing analyses, however, do not provide a reliable assessment of the typical degree of observer bias in randomised clinical trials. Thus, it is not clear whether observer bias in clinical trials, on average, is negligible or large or how variable the size and direction of any observer bias is or which factors in a trial are associated with a more pronounced degree of bias.

A reliable evaluation of the impact of non-blinded outcome assessors in randomised clinical trials is important, both to guide the design of future trials and to assist the balanced interpretation of trial results—for example, in the assessment of the risk of bias in trials for meta-analysis. It also seems important for evidence based medicine to strengthen its own evidence base. We systematically reviewed randomised trials with blinded and non-blinded assessors of binary outcomes to evaluate the impact of non-blinded outcome assessment on estimated treatment effects in randomised clinical trials and to examine reasons for its variation.

Methods

We included randomised clinical trials with blinded and non-blinded assessment of the same binary outcome. We excluded trials where it was unclear which group was experimental and which was control as such trials would not allow us to determine the direction of any bias; trials in which only a subgroup of patients had been evaluated by blinded and non-blinded assessors, unless they were selected at random; trials in which blinded and non-blinded assessors had access to each other’s results (for example, blinded assessments were provided to non-blinded assessors as a quality enhancement procedure); and trials where initially blinded assessors clearly had become unblinded—for example, when radiographs showed ceramic material indicative of the experimental intervention. Finally, we excluded trials with blinded end point committees adjudicating the assessments made by non-blinded clinicians because such adjudication often involves previous knowledge of the non-blinded assessment or is restricted to adjudication of events only.

We searched standard databases (PubMed, Embase, PsycINFO, CINAHL, Cochrane Central Register of Controlled Trials) and full text databases (HighWire Press and Google Scholar). Our core search string was: random* AND (“blind*” or unblind*) OR “masked and unmasked” with variations according to the specific database (see appendix on bmj.com). The last search was performed on 26 January 2010. We read the references of all included trials and asked authors of all the included trials if they knew of other trials.

One author (ASST) read all abstracts from standard databases and all text fragments from full text databases. If a study was potentially eligible, one author (ASST or AH) retrieved the full study report and excluded ineligible studies. Two authors (AH and ASST, SB, or BT) decided on the eligibility of the remaining studies. Disagreements were resolved by discussion.

We selected one binary outcome from each trial. If several outcomes had been assessed by both blinded and non-blinded assessors we selected the primary outcome of the trial, and if none was stated we selected the outcome we found most clinically relevant. We included the first assessment after the end of treatment, unless the primary outcome prescribed a different time point. Two authors (AH and either SB or BT) selected the outcome independently. Disagreements were resolved by discussion. For trials with more than two groups, we pooled the results in the experimental or the control groups.

We extracted background data for each trial (ASST and FE or AH and SB) and outcome data from each trial (AH and SB or BT): total number of failures and total number of successes in each group resulting from the blinded assessment and the non-blinded assessment. When possible we also extracted paired patient level data on blinded and non-blinded assessments, and constructed a 2x2 table (failure/success×blind/non-blind) for the experimental group and a corresponding table for the control group. Data from split body designed trials were treated as if they derived from parallel group trials.

If data were incomplete, we emailed the corresponding author and, if necessary, at least one additional author, followed up by telephone calls, and at least two reminders. Authors were asked whether they would share unpublished data with our group. We also searched the Food and Drug Administration (FDA) website for such data.

When authors chose to send us individual patient data (that is, all randomised patients listed by allocation group and result of blinded and non-blinded assessment), we checked whether all randomised patients were included in the dataset and tried to replicate a table or a main result of the published paper. Two authors (AH and BT or SB) independently derived outcome data. Any discrepancy was solved by discussion. We sent our results to the authors of the trial for comments.

For each trial, we evaluated five prespecified potential confounders in the comparison between blinded and non-blinded outcome assessments: a considerable time difference between these two assessments, different types of assessors (such as nurses v physicians), different types of procedures (such as direct visual assessment of wounds v assessment of photographs of wound), a substantial risk of ineffective blinding procedure, and non-identical groups of patients assessed (such as a few patients evaluated only by the blinded outcome assessor). For 16 trials, two masked authors (IB and PR) independently evaluated the first four items at a different location from the rest of the group. Other masked authors (AH and BT or SB) scored five trials. Disagreements were resolved by discussion. The masking was implemented by manipulating pdfs of the trial reports so that tables, graphs, or text describing results of any comparison between blinded and non-blinded assessors were blanked out. There were no cases of accidental unmasking.

Using the same masking procedure, we also evaluated characteristics of each outcome assessment. Two authors (mainly IB and PR) independently scored three factors out of a score of 5 (1 was low and 5 high): the degree of outcome subjectivity (that is, the degree of assessor judgment, high in assessment of...
global improvement and low in reading a laboratory sheet); the non-blinded outcome assessor’s overall involvement in the trial (that is, a proxy for the degree of personal preference for a result favourable to the experimental intervention); and the vulnerability of the outcome to the reporting and behaviour of non-blinded patients (as they might influence results considerably when outcomes are based on interviews and less so when outcomes are based on pure observations, such as inspection of radiographs). Disagreements were resolved by discussion.

We calculated the odds ratio for failures (such as an unhealed wound) in each trial for both the blinded and non-blinded assessments. An odds ratio under 1 indicates a beneficial effect of the experimental intervention. For each trial we summarised the impact of non-blinded outcome assessment as the ratio of the odds ratios (OR_{blind} / OR_{non-blind}). A ratio < 1 indicates that non-blinded assessments are more optimistic.

We meta-analysed the individual trial ratio of odds ratios with inverse variance methods using random-effects models. The standard error of the ratio of odds ratios used for the main analysis disregarded the dependency between blinded and non-blinded assessments. The statistical software we used was Stata 11.

We tested the robustness of our main analysis of the ratios of the odds ratios in sensitivity analyses. We used standard errors that took account of the dependence between blinded and non-blinded assessments (see appendix on bmj.com); all trials were given equal weight; and an analysis was conducted on the basis of the ratio of risk ratios, as risk ratios might be more easily interpretable than odds ratios by some. We studied whether the effect differed in subgroups of trials involving various types of data; clinical problems; objectives, designs, and sources of funding; and type of non-blinded assessor; and according to risk of confounding. We also evaluated the influence of small sample size on estimated ratio of risk ratios by funnel plot inspection.

We furthermore explored whether the variation in ratio of odds ratios was associated with the three prespecified outcome characteristic described above by random effects meta-regression of log ratio with the scores for each outcome characteristic.

To analyse the pattern of misclassifications underlying any difference between the blinded and non-blinded outcome assessments we compared the total number of failure events during non-blinded and blinded assessments in the experimental and in the control group and also compared the rate of agreement between blinded and non-blinded assessments in each trial. Finally, we calculated how many reclassifications of non-blinded assessments were needed to neutralise a difference between the blinded and non-blinded treatment effects—that is, to drive the ratio of odds ratios to 1 (see appendix on bmj.com).

**Results**

We examined 537 publications based on 1835 hits in standard databases and 2200 hits in full text databases. We excluded 512 studies, mostly because they were not randomised clinical trials or lacked blinded or non-blinded outcome assessment (see appendix on bmj.com). Thus, we included 25 trials (tables 1⇓ and 2). Of the 25 trials, six published outcomes for both types of assessments usable for our analysis. Contact with authors and searches of the FDA website increased the number of trials with outcome data to 21 (4391 randomised patients), of which eight trials provided individual patient data. Thirteen trials had strictly paired data (all patients had been assessed both by blinded and by non-blinded assessors), and eight trials provided predominantly paired data as a minority of patients had been assessed by only one type of assessor (see appendix on bmj.com).

In ten trials the validity of the non-blinded assessments was tested against the blinded assessments or non-blinded assessments were used as backup for missing blinded data. In four trials the main focus of the paper or abstract was a direct comparison between blinded and non-blinded outcome assessment, but it is unclear whether this was the original reason for using dual type assessors. In one trial refinement of the methods implied addition of blinded assessments without omission of the initially planned non-blinded assessments.

Fifteen of the 21 trials (71%) studied the effect of surgery or a procedure, 19 were parallel group trials (90%), and the median sample size was 172 (10th-90th centile 35-368). The trials were conducted in general surgery, orthopaedic surgery, plastic surgery, cardiology, gynaecology, anaesthesiology, neurology, psychiatry, dermatology, otorhynolaryngology, infectious diseases, and ophthalmology (table 1). The outcomes of the trials were in most cases subjective—for example, qualitative assessments of patients’ function (such as severity of angina or neurological deficit) or assessment of healing status (such as wounds or ulcers or fractures) (table 2). Seventeen trials (81%) scored 4 or 5 for outcome subjectivity on the 1 to 5 scale.

The odds ratio point estimate was more optimistic when based on the non-blinded assessors in 15 trials (71%) (fig 1). The ratio of odds ratios in the 21 trials ranged from 0.02 to 14.4 (fig 2). The pooled ratio of odds ratios was 0.64 (95% confidence interval 0.43 to 0.96) with moderate heterogeneity (I²=45%, P=0.015). Thus, on average, the odds ratios based on non-blinded assessments were exaggerated by 36% compared with the odds ratios based on blinded assessments.

Individual patient data provided 48% of the weight of the main analysis. The main result was robust, though sensitivity and subgroup analyses in general had wide confidence intervals (table 3). In the 12 trials with data on the dependence between blinded and non-blinded assessments the pooled ratio of odds ratios was 0.76 (0.61 to 0.94). In these 12 trials, the standard error accounting for the dependence was a median of 25% smaller than the corresponding standard errors assuming independence. Reducing the standard errors of the nine additional trials (without data on the dependence between blinded and non-blinded assessments) by 25% resulted in a pooled ratio of odds ratios of 0.64 (0.44 to 0.93). No trial was free from any of the five predefined possible confounders, but results were not clearly affected (table 3). The funnel plot was symmetrical on visual inspection (data not shown). Based on a qualitative assessment, the results in the four trials with incomplete or unclear outcome data did not to differ from the results in the trials we did meta-analyse (see appendix on bmj.com).

Meta-regression analyses showed no significant association between low ratios of odds ratios and scores for outcome subjectivity (P=0.27), non-blinded outcome assessor’s overall involvement in the trial (P=0.60), or outcome vulnerability to the reporting and behaviour of non-blinded patients (P=0.52). The slope of the regression line between log ratio of odds ratios and scores for outcome subjectivity, however, was in the expected direction. The 17 trials with clearly subjective outcomes (scores 4-5 on a 1-5 scale) had a pooled ratio of odds ratios of 0.55 (0.32 to 0.95). The five trials with moderately...
subjective outcomes (scores 2–3) had a pooled ratio of 0.93 (0.56 to 1.54).

The pattern of misclassifications underlying the difference between blinded and non-blinded results was characterised by more optimistic non-blinded assessments. The non-blinded assessors detected 26% fewer failure events (such as no wound healing) compared with the blinded assessors (984 vs 1335). In the intervention groups the non-blinded assessors detected 35% fewer patients with treatment failure than the blinded assessors (421 vs 649 events), whereas in the control group the proportion was 18% (563 vs 686 events).

The pattern of misclassifications was also characterised by a preoccupation with the intervention group. In the 12 trials with data on agreement, the blinded and non-blinded assessors agreed in a median of 78% of patient assessments (interquartile range 63–91%). The proportion of concordant assessments, and the corresponding proportion of discordant assessments, however, seemed to differ according to the allocation group. The median proportion of discordant assessments between blinded and non-blinded assessors per trial was 28% (9–41%) in the intervention group and 16% (9–37%) in the control group (see appendix on bmj.com).

The number of reclassified assessments per trial needed to neutralise a difference between the estimated blinded and non-blinded treatment effects (that is, to drive the ratio of odds ratios to 1.00) ranged from 0 to 41.7, with a median of 2.5. This corresponded to 0.28% of the assessed patients per trial, with a median of 3% (see appendix on bmj.com).

Discussion

The estimated effects of experimental interventions in randomised clinical trials tended to be considerably more optimistic when they were based on non-blinded assessment of subjective outcomes compared with blinded assessment. The pooled ratio of odds ratios was 0.64 (0.43 to 0.96), indicating that the non-blinded outcome assessors generated odds ratios that, on average, were exaggerated by 36%. We interpret this as empirical evidence for substantial observer bias.

Strengths and weaknesses of the study

This result is based on contemporary trials representing a fair range of clinical specialties. The unique trial design with paired data implies a low risk of confounding. The data were high quality, as individual patient data provided about half of the weight of the main analysis. Our results were robust to modifications to both type of analysis and summary statistic. For example, the ratio of relative risks was 0.78 (0.63 to 0.96), indicating that non-blinded outcome assessors generated relative risks that, on average, were exaggerated by 22%.

We possibly did not identify all trials but we do not know whether they would report markedly different results. Publication bias is normally driven by the effect of a treatment and has less impact on our comparison between two types of assessments. Four trials in our study published papers with a main focus on observer bias. Though confidence intervals were wide, these four trials did not report significantly different findings compared with the 17 other trials.

Our cohort of trials is not representative of medical trials in general. We included no trials with clearly objective outcomes, such as total mortality. The trials we did include had mainly subjective outcomes—such as qualitative assessments of patients and evaluation of fracture or wound healing—and our result is applicable to trials with similar subjective outcomes. We would anticipate less observer bias with more objective outcomes, though it is an interesting question which medical outcomes should be considered clearly objective, apart from total mortality and some laboratory outcomes. Furthermore, the extrapolation of our results to randomised trials with binary subjective outcomes hinges on the assumption that the degree of observer bias in our trials with dual observation of outcomes is essentially similar to trials with only non-blinded observers.

We found no association between observer bias and five prespecified potential confounders. A special concern, however, is consensus classifications that could reduce observer variability and leave less room for observer bias. The only trial with consensus based non-blinded assessments found no observer bias (ratio of odds ratios 1.06, 0.79 to 1.43). It is unclear whether this is caused by the consensus classification, chance, or other trial characteristics.

We included one trial with probable reversed direction of bias. The trial compared an experimental oral prodrg of ganciclovir, for cytomegalovirus retinitis with the intravenous version of the same substance, ganciclovir. The comparison between non-blinded and blinded outcome results in a ratio of odds ratios that was extreme, but in the reversed direction. Comparable retinitis trials, also with blinded and non-blinded assessors, have reported similar results favouring the control intervention on time to event outcomes.

Several previous studies have compared treatment effects in “double blind” trials with similar trials not reported as “double blind.” An overview of seven such studies reported a pooled ratio of odds ratios of only 0.91 (0.83 to 1.00). Wood and colleagues’ reanalysis of three of the studies reported a similar overall result but with a ratio of odds ratios of 0.75 (0.61 to 0.93) for subjective outcomes. These studies do not directly evaluate the impact of blinded outcome assessors, are partly based on ambiguous terminology, and involve a considerable risk of confounding. Still, our findings are numerically roughly similar to those of Wood and colleagues.

Mechanisms of observer bias

The pattern of misclassifications underlying the observer bias can be characterised by “optimism error” and “intervention preoccupation.” The non-blinded assessors detected fewer failures than blinded assessors. This optimism error, however, was much more pronounced in the intervention group than in the control group. Thus, the non-blinded outcome assessor did not “under-rate” patients in the control group and “over-rate” patients in the intervention group. Both groups were over-rated but the intervention group considerably more so.

A third important feature of observer bias is the striking contrast between the substantial degree of observer bias we found and the surprisingly small number of misclassified patients needed to generate this bias. The median number of patients needed to reclassify to neutralise bias in a trial was 2.5 or 3% of the assessed patients. The difference between numbers of events in the experimental group and the control group determines the estimated effect. Numbers of events are usually considerably smaller than the number of included patients, and still smaller is the number of misclassifications needed to bias the estimated effect. For example, in the trial by Noseworthy and colleagues, the ratio of odds ratios was 0.81 (0.40 to 1.61). This degree of bias was neutralised by reclassification of two
of the 140 included patients. Binary outcomes seem sensitive to directional misclassifications of a few patients.

Fundamentally, observer bias is caused by the predispositions of the observers, which might vary unpredictably from trial to trial. Our cohort of trials probably consists of some trials with largely neutral assessors and some trials with predisposed assessors. The expected degree of observer bias in trials with predisposed assessors will be considerably larger than our averaged result. Thus, in any individual trial it is not possible to safely predict neither the direction nor the size of any bias. We would merely pool the data as a simplistic correction factor. When the possible bias in a trial with non-blindedy assessors is ascertained, the range of possible observer bias should be taken into account and not only our pooled average. Furthermore, it would be prudent to also consider the type of outcome involved and any indicators for predispositions in assessors.

Implications

Blinding outcome assessors might be seen as too cumbersome, unnecessary, or directly mistaken; 37 compared with the huge logistical challenges involved in setting up a trial, however, it is a minor procedure and one that improves reliability considerably. Fortunately, blinding the assessor is possible in nearly all trials, sometimes after the development of creative blinding procedures. 40, 41 In some trials a subsample of patients is blindly assessed and the result used to validate non-blindedy assessments. Such comparisons are inherently underpowered and should be avoided.

Our result strengthens the hypothesis that blinding can also be important for other key people in a trial, especially patients, 42 who can be seen as privileged outcome assessors of their own symptoms. Still, it is important to separately study the impact of blinding each key person. For example, one study found little impact of blinded outcome adjudicators in 10 large cardiovascular trials. 43

We found no significant association between the degree of observer bias and degree of outcome subjectivity, though the association was in the expected direction. Future investigations could further analyse the role of outcome subjectivity and other factors that could modify the degree of observer bias.

The problem of observer bias goes beyond the randomised clinical trial. Comparisons between blinded and non-blindedy observers in other types of empirical investigations have reported results indicative of observer bias—for example, in an observational study of patients with primary dystonia, 44 an evaluation of cancer staging, 45 an assessment of surgical skills, 46 and a neurophysiological laboratory study. 47 Furthermore, observer bias has been reported or discussed within veterinary science, 48 forensic science, 49 special education studies, 50 animal behaviour research, 51 and broadly within psychology. 52

Observation is fundamental to scientific activity; observer bias might be too.

In conclusion, randomised clinical trials without blindedy observers of subjective binary outcomes will, on average, generate substantiy biased estimates of treatment effects. The bias is compatible with a high rate of agreement between blinded and non-blindedy assessments and is driven by the misclassification of a few patients.

We thank the following researchers for sharing unpublished outcome data with us: Andrew Jull and the Clinical Trials Research Unit (CTRU) at the University of Auckland, Alexandre Valentin-Opran, Nicky Cullum, Tim Reynolds, Peggy Vandervoort, and George C Ebers (individual patient data); and Daniel Burkhoff, Amy P Murtha, and Cheryl Iglesia (detailed outcome data). We also thank Peter C Gatzsche for valuable comments on previous versions of the manuscript.

Contributors: AH conceived the idea and design, organised the study, and wrote the first draft of the manuscript. ASST and AH developed the search strategy. ASST, FE, BT, SB, and AH did the non-blinded data collection. IB and PR did the masked data collection, supplemented by SB, BT, and AH. AH and JH did the statistical analyses. All authors discussed the result and commented on the manuscript. AH is guarantor.

Funding: The study was partially funded by the Danish Council for Independent Research: Medical sciences. The funder had no influence on the study design and the collection of data, the writing and the article to decide it to publication.

Competing interests: All authors have completed the ICMJE uniform 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 organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: Not required.

Data sharing: Statistical code and dataset available from the corresponding author.

1 Higgins JPT, Green S, eds. Cochrane handbook for systematic reviews of interventions. Version 5.0.0. Cochrane Collaboration, 2008.

2 Rosenthal R. Experimenter effects in behavioral research. Appleton-Century-Crofts, 1966:13-4.

3 Karanikolas PJ, Bhandari M, Tamhane B, Aki EA, Bassler D, Alonso-Coello P, et al. Blinding of outcomes in trials of orthopedic trauma: an opportunity to enhance the validity of clinical trials. J Bone Joint Surg Am 2008;90:1029-33.

4 Haahr M, Heldropatanon A. Who is blind in randomised clinical trials? An analysis of 200 trials and a survey of authors. Clin Trials 2006;3:360-5.

5 Noseworthy JH, Ebers GC, Vandervoort MK, Farquhar RE, Yeltis E, Roberts R. The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. Neurology 1994;44:16-20.

6 Postman RW, Struja PA, Kripa R, Sirovetz IN, Mark RW, Faroohi Y, et al. Reporting of outcomes in orthopedic randomized trials: does blinding of outcome assessors matter? J Bone Joint Surg Am 2007;89:559-8.

7 Pitlaid J, Heldropatanon A, Jergensen KJ, Hilden J, Altman DG, Gøtzsche PC. Impact of allocation concealment on conclusions drawn from meta-analyses of randomised trials. Int J Epidemiol 2007;36:847-57.

8 Wood L, Egger M, Glasziou PL, Schütz FH, Jüni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ 2008;336:601-5.

9 DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986:7:177-88.

10 Byrkhoff D, Schmidt S, Schulman SP, Myers J, Resar J, Becker LC, et al. Transmyocardial laser revascularisation compared with continued medical treatment for refractory angina pectoris: a prospective randomised trial. ATLANTIC Investigators. Angina Treatments-Lasers and Normal Therapies in Comparison. Lancet 1999:354:885-90.

11 Dumville JC, Worthy G, Bland JM, Cullum N, Dowson C, Iglesias C, et al. Lavlar therapy for leg ulcers (Viens APR): a randomised controlled trial. BMJ 2009;338:b773.

12 Gouwender S, Clamín C, Gerant HK, Valentin-Opran A, Amt Y, Aresel R, et al. Recombinant human bone morphogenetic protein-2 for treatment of open full-thickness fractures: a prospective, controlled, randomized study of four hundred and fifty patients. J Bone Joint Surg Am 2002:84-A:2123-34.

13 Jones AL, Bucholz RW, Bosse MJ, Mirza SK, Lyon TR, Webb LX, et al. Recombinant human bone morphogenetic protein-2 to prevent development of open full-thickness fractures: a prospective, randomized trial in open tibial fractures treated with reamed nailing. J Bone Joint Surg Am 2011:93:801-8.

14 Aro HT, Govender S, Patel AD, Hernigou P, Perera de Gregorio A, Popescu GI, et al. Recombinant human bone morphogenetic protein-2: a randomized trial in open tibial fractures treated with reamed nailing. J Bone Joint Surg Am 2002:84-A:2123-34.

15 Jull A, Walker N, Parag V, Molan P, Rodgers A, for the Honey as Adjuvant Leg Ulcer Therapy Trial Collaborators. Randomized clinical trial of honey impregnated dressings for venous leg ulcers. Br J Surg 2009:96:175-82.

16 Landsmann AS, Robbins AH, Angelini PF, Wu CC, Cook J, Oster M, et al. Treatment of mild, moderate, and severe onychomycosis using 870- and 930-nm light illumination. J Am Podiatr Med Assoc 2010;100:166-77.

17 Martin DF, Sierra-Madero J, Walmays S, Woltz RA, Macay K, Georgiou P, et al. A controlled trial of valgancylic acid as induction therapy for cytogenetically resistant. N Engl J Med 2002:346:1119-26.

18 Meltzer HY, Alphei L, Green AI, Atamura AC, Anand R, Beredjik L, et al. Clozapine treatment for suicidality in schizophrenia: International Suicide Prevention Trial (InterSePT). Arch Gen Psychiatry 2000;57:802-9.

19 Miller RS, Steward DL, Tami TA, Sillars MJ, Seiden AM, Shete M, et al. The clinicaleffects of hyaluronic acid ester nasal dressing (Merogel) on intranasal wound healing after transnasal endoscopic sinus surgery. Otolaryngol Head Neck Surg 2003;128:166-77.

20 Murtha AP, Kaplan AL, Paglia MJ, Miles BB, Feldstein ML, Rutf GL. Evaluation of a novel technique for wound healing using a barrier sponge. Plast Reconstr Surg 2006;117:1769-80.

21 Noseworthy JH, Vandervoort MK, Penman M, Ebers G, Shumak K, Seland TP, et al. Cyclophosphamide and plasma exchange in multiple sclerosis. Lancet 1991:337:1540-1.
What is already known on this topic

Non-blinded assessors of binary outcomes are used in many randomised trials

It is prudent to suspect bias in randomised clinical trials with non-blinded outcome assessors

What this study adds

Estimated effects in randomised clinical trials, measured as odds ratios, are exaggerated by an average of 36% when based on non-blinded assessments of subjective binary outcomes

The bias is compatible with a high rate of agreement between blinded and non-blinded outcome assessors and driven by the miscategorisation of few patients
Table 1 | Characteristics of 25 included randomised clinical trials in study of effect of blinded or non-blinded outcome assessors

|                          | No (%) |
|--------------------------|--------|
| **General:**             |        |
| Parallel group design    | 19 (94) |
| Two study groups         | 17 (81) |
| Primary outcome defined  | 17 (81) |
| Intervention: surgery or procedure | 14 (67) |
| Intervention: drug       | 5 (24)  |
| Control group: standard care | 18 (86) |
| Control group: no treatment/placebo | 3 (14)  |
| Published in specialty journal (such as Annals of Surgery) | 15 (75) |
| Published in general medical journal (such as Lancet) | 5 (25)  |
| **Outcome:**             |        |
| Clearly subjective (score 4-5 on 1-5 scale) | 16 (76) |
| Moderately subjective (score 2-3) | 5 (24)  |
| Objective (score 1)      | 0      |
| **Medical specialty:**   |        |
| Cosmetic surgery         | 4 (19)  |
| General surgery          | 4 (19)  |
| Orthopaedic surgery      | 3 (14)  |
| Dermatology/ophthalmology/otolaryngology | 3 (14)  |
| Cardiology               | 2 (10)  |
| Neurology/psychiatry     | 2 (10)  |
| Gynaecology              | 1 (5)   |
| Anaesthesia              | 1 (5)   |
| Infectious diseases      | 1 (5)   |
| **Trial methods:**       |        |
| Random allocation sequence adequately generated | 6 (29)  |
| Random allocation sequence adequately concealed | 12 (57) |
| Patients blinded         | 8 (39)  |
| Treatment provider blinded | 0      |
| Drop outs accounted for (ITT analysis or no drop outs) | 13 (62) |

ITT=intention to treat.
| Trial                          | No of patients | Clinical problem                  | Experimental v control                       | Outcome                                      | Assessment                                      | Blind vs Non-blinded |
|-------------------------------|----------------|-----------------------------------|----------------------------------------------|----------------------------------------------|------------------------------------------------|----------------------|
| Smith 2007                   | 117            | Facial folds                      | Hydroxylaplatit v collagen                   | No improvement; GAIS, 6 months               | Photo of folds, 3 evaluators                    | Live inspection of folds, clinician       |
| MA-1300-15                    | 180            | Thin lips                         | Hyaluronic acid v no treatment               | No improvement (±1) MLFS, week 12            | Inspection of lips, evaluator                  | Inspection of lips, clinician             |
| Oesterle 2000                 | 221            | Angina pectoris                   | Laser (TMR) v drugs only                    | CCSA class III/IV, 12 months                | Interview by assistant, CCSA grade: cardiologist | Interview and CCSA grade: cardiologist |
| Meltzer 2003                  | 980            | Suicide risk                      | Olanzapine v clozapine                      | Worse/much worse; CGI-SS, 24 months         | Clinical assessment, psychiatrists              | Clinical assessment, psychiatrists         |
| Landsman 2010                 | 36             | Onychomycosis                      | Light therapy v sham light                   | Not markedly improved; day 180              | Photo of nail, expert panel                    | Live nail inspection, clinicians          |
| Burkoff 1999                  | 182            | Angina pectoris                   | Laser (TMR) v drugs only                    | CCSA class III/IV, 12 months                | Interview by assistant, CCSA grade: cardiologist | Interview and CCSA grade: cardiologist |
| Reynolds 2004                 | 35             | Wound                             | Vacutex v standard dressing                 | No improvement; day 29                      | Wound photo, 9 nurses                          | Live inspection, nurse                    |
| Jones 2006                   | 30             | Fracture                          | rhBMP-2 allograft v autogenous graft         | No fracture union, 6 months                 | Radiograph, radiologist                       | Radiograph, clinician                     |
| Aro 2011                     | 277            | Fracture                          | rhBMP-2 v standard care only                | No fracture union, 13 weeks                 | Radiograph, radiologist                       | Radiograph, clinician                     |
| Miller 2003                   | 50             | Nasal wound                       | MeroGel v Merocel dressing                  | Synechia, last follow-up                    | Endoscopic image, 3 investigators              | Live endoscopy, clinician                 |
| Reynolds 2004                 | 142            | Wound                             | Drawtex v standard dressing                 | No improvement, day 29                      | Wound photo, 9 nurses                          | Live inspection, nurse                    |
| Jull 2008                    | 368            | Leg ulcers                        | Honey dressing v usual dressing             | Unhealed ulcers, week 12                    | Ulcer photo, reviewer                         | Live ulcer inspection, nurses             |
| Noseworthy 1994              | 168            | Multiple sclerosis                | Plasma exchange v CPM v placebo             | EDSS score increase ≥1, 12 months           | Examination, neurologist                      | Examination, neurologist                 |
| Brandstrup 2003               | 172            | Fluid regimens                    | Restricted v standard                       | Postoperative complications, day 30         | Censured medical records; 2 surgeons          | Standard medical records, surgeon         |
| Wabel 2006                   | 20             | Vaccine site                      | Upper inner arm v outer defloid             | Non-take*, day 7                            | Skin photo, investigator                      | Live skin inspection, investigator        |
| Dumville 2009                | 267            | Venous ulcers                     | Larvae v wound dressing v standard suture   | Not healed ulcers, week 26                  | Ulcer photo, 2 assessors                      | Live ulcer inspection, 2 nurses           |
| Munha 2006                   | 188            | Scar cosmesis                     | Barbed suture v standard suture             | Modified HCS (≥2), week 5                   | Scar photo, surgeons                          | Live scar inspection, clinicians          |
| Iglesia 2010                  | 65             | Vaginal prolapse                  | Polypropylene mesh v standard               | Prolapse recurrence; POP-Q >1, 3 months     | Examination, several evaluators (such as nurse or urogyneacologist) | Examination, surgeon                      |
| Dover 2009                   | 283            | Facial folds                      | Large particle hyaluronic acid v small       | No improvement; WSR5, week 12               | Inspection of facial folds, evaluator          | Inspection of facial folds, clinician      |
| Martin 2002                  | 160            | CMV retinitis                     | Valganciclovir v ganciclovir                | Progression of retinitis, 4 weeks           | Fundus photo, evaluator                       | Live ophthalmoscop inspection, clinician  |

GAIS = global aesthetic improvement scale; MLFS = Medicis lip fullness scale; CCSA= (grading of) pectoris; CGI-SS=clinical global impression on suicide severity scale (7 point version); TMR=transmyocardial laser revascularisation; rhBMP-2=recombinant human bone morphogenetic protein-2; CPM=cyclophosphamide; EDSS=expanded disability status scale; HCS=Hollanders cosmetic score; POP-Q=pelvic organ prolapse quantification exam; WSR5=wrinkle severity rating scale; CMV=cytomegalovirus.

*Such as lack of pustular lesion.
| Comparisons                                                                 | No of trials | I² (P value) | ROR (95% CI) |
|----------------------------------------------------------------------------|--------------|--------------|--------------|
| Main analysis                                                              | 21           | 45% (0.02)   | 0.64 (0.43 to 0.96) |
| Type of analysis:                                                           |              |              |              |
| Dependence between blinded and non-blinded assessments accounted for*:     |              |              |              |
| Paired (blinded, non-blinded) patient level data                           | 12           | 11% (0.34)   | 0.76 (0.61 to 0.94) |
| As above plus correction factor                                              | 21           | 70% (<0.001) | 0.64 (0.44 to 0.93) |
| All trials given same weight                                                | 21           | NA           | 0.60 (NA)    |
| Direction of bias reversed in one trial (Martin 2002¹)                      | 21           | 40% (0.03)   | 0.57 (0.39 to 0.84) |
| Type of data:                                                               |              |              |              |
| Individual patient data (IPD)                                              | 8            | 0% (0.99)    | 0.77 (0.55 to 1.09) |
| Paired patient level data (blinded, non-blinded), no IPD                   | 4            | 0% (0.69)    | 0.38 (0.17 to 0.81) |
| Summary outcome (no paired patient level data)                             | 9            | 74% (<0.001) | 0.57 (0.19 to 1.69) |
| Clinical problem:                                                           |              |              |              |
| Wound/ulcer                                                                | 5            | 0% (0.96)    | 0.84 (0.55 to 1.29) |
| Fractured bone                                                             | 3            | 0% (0.98)    | 0.63 (0.33 to 1.20) |
| Angina pectoris                                                            | 2            | 0% (0.73)    | 0.31 (0.13 to 0.71) |
| Facial folds                                                               | 2            | 95% (<0.001) | 0.22 (0.00 to 19.8) |
| Other problems only studied in one trial                                    | 9            | 35% (0.14)   | 0.78 (0.37 to 1.64) |
| Trial characteristics:                                                      |              |              |              |
| Non-blinded assessment:                                                     |              |              |              |
| Multiple observer consensus                                                 | 1            | NA           | 1.06 (0.48 to 2.37) |
| Single observer                                                            | 20           | 46% (0.02)   | 0.61 (0.40 to 0.94) |
| Publication status:                                                         |              |              |              |
| Observer bias main objective                                                | 4            | 0% (0.92)    | 0.84 (0.41 to 1.71) |
| Observer bias not main objective                                           | 17           | 55% (0.004)  | 0.59 (0.36 to 0.97) |
| Design:                                                                    |              |              |              |
| Parallel group design                                                       | 19           | 17% (0.25)   | 0.74 (0.54 to 1.02) |
| Split-body                                                                 | 2            | 82% (0.02)   | 0.12 (0.04 to 3.74) |
| Funding:                                                                   |              |              |              |
| Industry                                                                   | 14           | 60% (0.002)  | 0.51 (0.26 to 0.97) |
| Non-commercial source or unclear                                           | 7            | 0% (0.99)    | 0.88 (0.60 to 1.29) |
| Risk of confounding:                                                        |              |              |              |
| Timing of blinded and non-blinded assessment:                              |              |              |              |
| Same/similar                                                               | 21           | 45% (0.02)   | 0.64 (0.43 to 0.96) |
| Not same/similar                                                           | 0            | NA           | NA           |
| Assessors:                                                                 |              |              |              |
| Same type (such as neurologist v neurologist)                              | 10           | 62% (0.005)  | 0.50 (0.22 to 1.10) |
| Not same type (such as radiologist v surgeon)                              | 11           | 20% (0.26)   | 0.72 (0.48 to 1.07) |
| Procedure:                                                                 |              |              |              |
| Same type (such as radiographs v radiographs)                              | 5            | 56% (0.06)   | 0.72 (0.27 to 1.92) |
| Not same type (such as photo v live observation)                           | 16           | 43% (0.04)   | 0.61 (0.39 to 0.95) |
| Blinding procedures:                                                        |              |              |              |
| Probably effective                                                         | 15           | 46% (0.03)   | 0.81 (0.50 to 1.33) |
| Possibly not effective                                                     | 6            | 0% (0.49)    | 0.40 (0.24 to 0.68) |
| Patients:                                                                  |              |              |              |
| All seen by both assessors                                                 | 13           | 0% (0.91)    | 0.70 (0.52 to 0.96) |
| Minority seen by only one type of assessor                                 | 7            | 77% (<0.001) | 0.50 (0.15 to 1.76) |

NA=not assessable or no data. *Standard error in 12 trials with data on dependence between blinded and non-blinded assessments (paired patient level data) reduced by median of 25% when dependence was incorporated. This median reduction was used as correction factor in 9 trials without such data.
Figures

Fig 1 Estimated intervention effect according to blinded or non-blinded outcome assessor
Fig 2 Impact of non-blinded outcome assessors on estimated intervention effects in randomised clinical trials measured as ratio of odds ratios (odds ratio based on non-blinded outcome assessors divided by odds ratio based on blinded outcome assessors)