Observer bias in randomized clinical trials with measurement scale outcomes: a systematic review of trials with both blinded and nonblinded assessors

Asbjørn Hróbjartsson MD PhD, Ann Sofia Skou Thomsen MD, Frida Emanuelsson MD, Britta Tendal MD PhD, Jørgen Hilden MD, Isabelle Boutron MD PhD, Philippe Ravaud MD PhD, Stig Brorson MD PhD

Abstract

Background: Clinical trials are commonly done without blinded outcome assessors despite the risk of bias. We wanted to evaluate the effect of nonblinded outcome assessment on estimated effects in randomized clinical trials with outcomes that involved subjective measurement scales.

Methods: We conducted a systematic review of randomized clinical trials with both blinded and nonblinded assessment of the same measurement scale outcome. We searched PubMed, EMBASE, PsycINFO, CINAHL, Cochrane Central Register of Controlled Trials, HighWire Press and Google Scholar for relevant studies. Two investigators agreed on the inclusion of trials and the outcome scale. For each trial, we calculated the difference in effect size (i.e., standardized mean difference between nonblinded and blinded assessments). A difference in effect size of less than 0 suggested that nonblinded assessors generated more optimistic estimates of effect. We pooled the differences in effect size using inverse variance random-effects meta-analysis and used metaregression to identify potential reasons for variation.

Results: We included 24 trials in our review. The main meta-analysis included 16 trials (involving 2854 patients) with subjective outcomes. The estimated treatment effect was more beneficial when based on nonblinded assessors (pooled difference in effect size –0.23 [95% confidence interval (CI) –0.40 to –0.06]). In relative terms, nonblinded assessors exaggerated the pooled effect size by 68% (95% CI 14% to 230%). Heterogeneity was moderate (I^2 = 46%, p = 0.02) and unexplained by metaregression.

Interpretation: We provide empirical evidence for observer bias in randomized clinical trials with subjective measurement scale outcomes. A failure to blind assessors of outcomes in such trials results in a high risk of substantial bias.

Competing interests: Frida Emanuelsson and Ann Sofia Skou Thomsen have received grants from the Danish Council of Independent Research. No other competing interests were declared.

This article has been peer reviewed.

Correspondence to:
Asbjørn Hróbjartsson,
ah@cochrane.dk

***CMAJ* 2013. DOI:10.1503/cmaj.120744**

A failure to blind assessors of outcomes in randomized clinical trials may result in bias. Observer bias, sometimes called "detection bias" or "ascertainment bias," occurs when outcome assessments are systematically influenced by the assessors' conscious or unconscious predispositions — for example, because of hope or expectations, often favouring the experimental intervention.[1]

Blinded outcome assessors are used in many trials to avoid such bias. However, the use of nonblinded assessors remains common,[2-4] especially in nonpharmacological trials; for example, nonblinded outcome assessment was used in 90% of trials involving orthopedic traumatology[3] and 74% of trials involving strength training for muscles.[4]

Unfortunately, the empirical evidence on observer bias in randomized clinical trials has been incomplete. Meta-epidemiological studies have compared double-blind trials with similar trials that were not double-blind.[5,6] However, such studies address blinding crudely because "double-blind" is an ambiguous term.[3,7] Furthermore, the risk of confounding is considerable in indirect between-trial analyses, as "double-blind" trials may have better overall methods and larger sample sizes than trials that are not reported as "double-blind."

A more reliable approach involves analyses of trials that use both blinded and nonblinded outcome assessors, because such a within-trial design provides a direct comparison between blinded and nonblinded assessments of the same outcome in the same patients. Our previous analysis of such trials with binary outcomes found substantial observer bias.[8]

Although subjective measurement scales such as illness severity scores are popular, they may be susceptible to observer bias. They are frequently used as outcomes in clinical scenarios with no naturally distinct categories, and adjacent subcategories on a scale typically involve minor and vaguely defined differences.

We decided to systematically review trials with both blinded and nonblinded assessment of outcomes using the same measurement scales. Our primary objective was to evaluate the impact of nonblinded outcome assessment on estimated treatment effects in randomized clinical trials. Our secondary objective was to examine reasons for variation in observer bias.

Methods

Eligibility criteria

We included randomized clinical trials with blinded and nonblinded assessment of the same measurement scale outcome. We excluded trials for which the distinction between the experimental and control groups was unclear, because such trials would not allow us to determine the direction of any bias; trials for which only a subgroup of patients were evaluated by blinded and nonblinded assessors, unless selected at random; trials in which blinded and nonblinded assessors had access to each others' results; and trials in which initially blinded assessors became unblinded (e.g., when radiographs showed ceramic material indicative of the experimental intervention).

Search strategy

We searched the following databases from their inception onwards without language restrictions: PubMed, EMBASE, PsycINFO, CINAHL, The Cochrane Central Register of Controlled Trials, HighWire Press and Google Scholar. Our core search string was random* AND ("blind* and unblind*" OR "masked and unmasked") with variations according to the specific database (Appendix 1, available at www.cmaj.ca/lookup/suppl/doi:10.1503/cmaj.120744/-/DC1). We performed the last search on Jan. 26, 2010. We read the references of all of the included trials and asked the authors of all included trials whether they knew of additional trials to identify any further studies that should be included.

Data abstraction

One investigator read all abstracts from standard databases and all text fragments from full-text databases. If a study was identified as potentially eligible for inclusion, we retrieved a full study report, which was read by an investigator who excluded all clearly ineligible studies. Two investigators read all other study reports and decided on eligibility. Disagreements were resolved by discussion.

We selected a single measurement scale from each trial. If several outcomes had been assessed under both blinded and nonblinded conditions, we preferred the primary outcome of the trial and the first assessment after the end of treatment (unless the primary outcome prescribed a different time point). Two investigators selected the outcomes independently. Again, disagreements were resolved by discussion. For trials with more than 2 groups, we pooled the results in the experimental or control groups.[1]

From each trial we extracted the following data: posttreatment mean, standard deviation and the numbers of patients in the experimental and control groups in the blinded assessments, and the corresponding data from the nonblinded assessments. For crossover and split-body trials, we extracted the standard deviation of the paired difference between treatments. If possible, we also extracted data on the correlation between blinded and nonblinded assessment (e.g., Spearman rank correlation coefficient) and data on interobserver variation between assessors (blinded or nonblinded).

If data were incomplete, we contacted the authors of the trial by email or telephone. We also searched the US Food and Drug Administration (FDA) website for trial outcome data. If standard deviations were not reported, we used standard deviations from a comparable trial that used the same measurement scale. If interobserver data were not available, we tried to obtain them from independent scale-validation studies.

For each trial, we evaluated 5 prespecified potential confounders in the comparison between blinded and nonblinded outcome assessments: a considerable time lapse between the 2 assessments, different types of assessors (e.g., nurses v. physicians), different assessment procedures (e.g., direct visual assessment of a wound v. a photograph of a wound), a substantial risk of ineffective blinding and different patients being assessed (i.e., some patients who had been evaluated blindly had not been evaluated nonblindly and vice versa). The first 4 items were evaluated by 2 investigators masked to any information relating to the comparison between blinded and nonblinded assessors. The masking was done by manipulating PDF versions of the trial reports so that tables, graphs or text describing the results of any comparison between blinded and nonblinded assessors were blanked out. There were no cases of accidental unmasking.

In addition, for each trial, we evaluated 3 characteristics of the outcomes that could possibly explain variations in observer bias. Two masked investigators independently evaluated the following 3 factors on a scale from 1 to 5 (1 = low, 5 = high): the degree of subjectivity of the outcome (i.e., the degree to which the assessors' judgment affected the outcome; high in global assessment of patient improvement and low in reading a laboratory sheet); the nonblinded assessor's overall involvement in the trial (i.e., a proxy for the degree of personal preference for a result favourable to the experimental intervention); and the vulnerability of the outcome to nonblinded patients (high in outcomes based on interviews with nonblinded patients and low in outcomes involving pure observation, such as the inspection of photographs). Disagreements were resolved by discussion.

Statistical analysis

For each trial, we calculated the effect size (i.e., standardized mean difference) based on the blinded and nonblinded assessments using the pooled standard deviation of the blinded assessments as the common standardizing unit. An

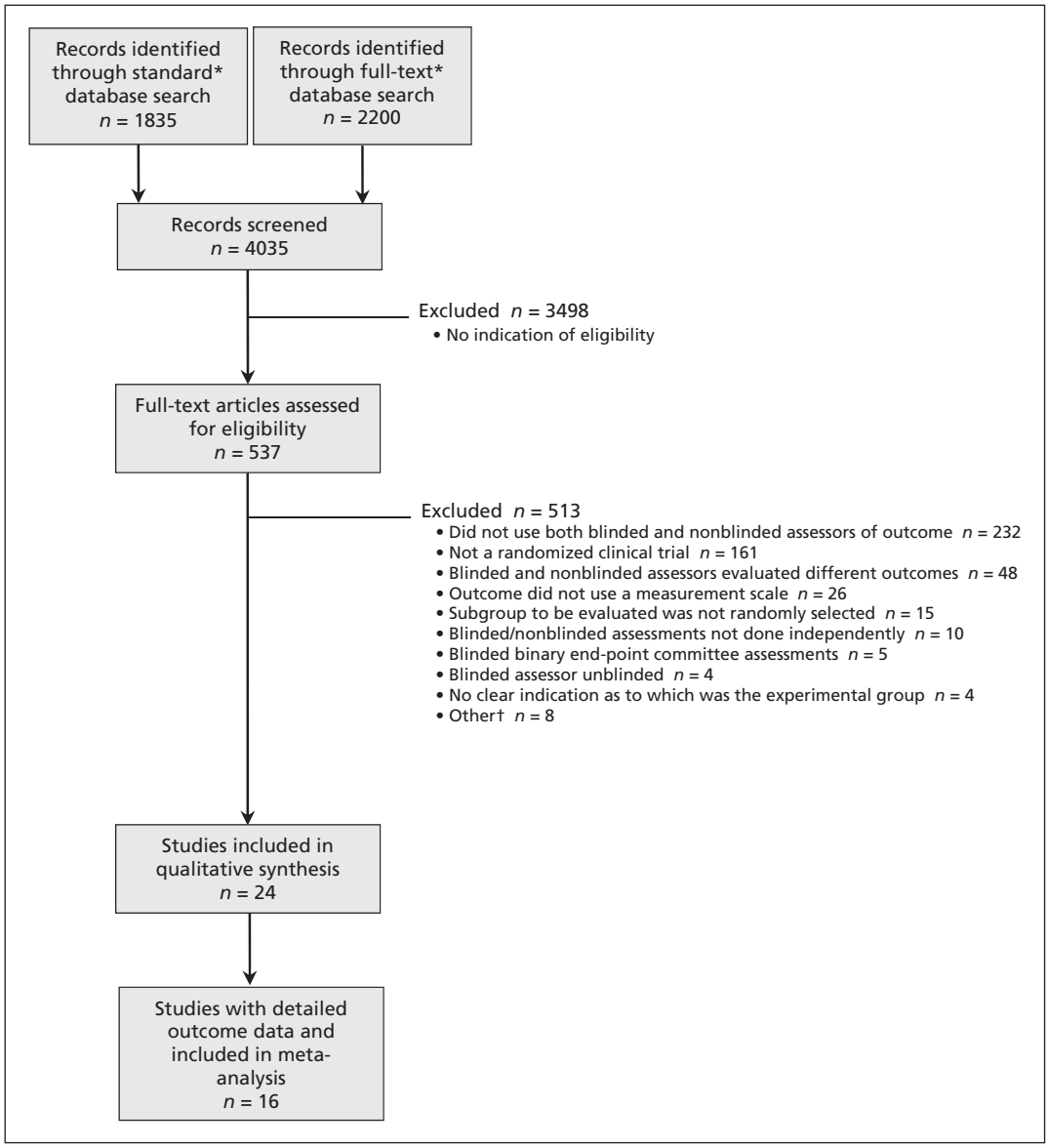

Figure 1: Flow diagram for identification of eligible trials. *A standard database (e.g., Medline) indexes publications that are searchable by title, keywords and abstract, but does not contain the full text of the publication; a full-text database (e.g., Google Scholar) indexes the searchable full-text of publications. †Other reasons for exclusion incude different interventions for patients assessed by blinded and nonblinded assessors, retrospective analysis of a risk factor, the same patients as involved in another included trial, lack of clarity as to whether nonblinded clinicians formally assessed outcomes or the use of blinded versus nonblinded assessment in only 1 arm of the trial.

effect size of less than 0 suggests a beneficial effect of the experimental intervention. We subsequently summarized the impact of nonblinded outcome assessment as the difference between the 2 effect sizes. A difference in effect size of less than 0 suggests that the nonblinded assessments generate more optimistic estimates of effect than do the blinded assessments.

We pooled the differences in effect size from individual trials by meta-analysis using random-effects models and inverse variance weights.[9] The standard error of the difference in effect size used for the main analysis disregarded the correlation between blinded and nonblinded assessments (Appendix 2, available at www.cmaj.ca/lookup/suppl/doi:10.1503/cmaj.120744/-/DC1).

We tested the robustness of our main analysis with secondary analyses addressing the type of analysis (e.g., incorporating the correlation between blinded and nonblinded assessments), type of data, clinical condition, trial characteristics, risk of confounding and trial size. In addition, we examined the percentage by which the nonblinded effect estimate exceeded the blinded effect estimate (effect size difference/blinded effect size), approximating the confidence interval for the percentage according to Fieller.[10]

Finally, we used univariable random-effects metaregression to determine whether variations in effect size differences were associated with the 3 prespecified outcome characteristics we described earlier.

Table 1: Characteristics of randomized clinical trials included in our meta-analysis

Characteristic	No. (%) $n = 16$
General	
Parallel group design	13 (81)
2 study groups	15 (94)
Primary outcome defined	13 (81)
Type of intervention	
Surgery/procedure	11 (69)
Drug	5 (31)
Control group	
Standard care	12 (75)
No treatment/placebo	4 (25)
Type of publication*	
Specialty journal	11 (73)
General medical journal	4 (27)
Funding source	
Industry	10 (63)
Noncommercial or unclear	6 (38)
Subjectivity of outcome (score on scale of 1–5)	
Clearly subjective (4–5)	13 (81)
Moderately subjective (2–3)	3 (19)
Objective (1)	0 (0)
Medical specialty	
Neurology	4 (25)
Cosmetic surgery	3 (19)
Cardiology	2 (13)
Psychiatry	2 (13)
Otolaryngology	2 (13)
Dermatology, gynecology or infectious diseases	3 (19)
Trial methods	
Random allocation sequence adequately generated	2 (13)
Random allocation sequence adequately concealed	6 (38)
Patients blinded	7 (44)
Treatment provider blinded	1 (6)
Drop-outs < 15%	8 (50)

*One trial was unpublished.

Results

We identified 537 publications from 1835 hits in standard databases and 2200 hits in full-text databases. We excluded 513 studies, mostly because they were not randomized clinical trials or because they lacked blinded or nonblinded outcome assessment (Figure 1). Thus, 24 trials were included in our qualitative synthesis.[11–36]

Of these 24 trials, 16 (involving 2854 patients) provided outcome data for both the blinded and nonblinded assessors. The characteristics of the trials are described in Table 1. The clinical specialties represented were neurology, cosmetic surgery, cardiology, psychiatry, otolaryngology, dermatology, gynecology and infectious diseases.

The outcomes of the trials were generally subjective; 13 of the 16 trials (81%) scored 4 or 5 on our scale of subjectivity (Table 2). The median Spearman rank correlation coefficient between blinded and nonblinded assessments in the 7 trials with such data was 0.67 (Appendix 3, available at www.cmaj.ca/lookup/suppl/doi:10.1503/cmaj.120744/-/DC1.). We identified validation studies for scales used in 10 of the included trials, which generally reported good interobserver agreement (median weighted κ 0.64 [5 trials]; median intraclass correlation coefficient 0.82 [5 trials] (Appendix 3).

In 10 trials (63%), the effect size point estimate was more optimistic as determined by the nonblinded assessors (Figure 2). Among all 16 trials, the difference in effect size ranged from –1.10 to 0.14. The pooled difference in effect size was –0.23 (95% confidence interval

Table 2: Characteristics of the outcome assessments in the trials included in our meta-analysis

Trial	No. of patients	Clinical condition	Experimental v. control	Outcome	Assessment	
					Blind	Nonblind
Cohen et al[11,12]	285	Facial wrinkles	Artecoll v. collagen	Facial fold assessment scale (0–5), 6 mo	Photos, 3 investigators	Inspection, treating investigator
Oesterle et al.[13]	221	Angina pectoris	Laser (TMR) v. medication only	CCSA class (I–IV), 12 mo	Interview, assistant; CCSA, cardiologist	Interview and CCSA, cardiologist
Powell et al.[14]	22	Turbinate hypertrophy	Radiofrequency reduction v. sham	Nasal obstruction (10-cm VAS), 4 wk	Nasal examination (rhinoscopy)	Nasal examination (rhinoscopy)
Burkhoff et al.[15]	182	Angina pectoris	Laser (TMR) v. medication only	CCSA class (I–IV), 12 mo	Interview; assistant; CCSA, cardiologist	Interview and CCSA, cardiologist
Wedekind et al.[16]	75	Panic disorder	Aerobic exercise v. relaxation	PAS (0–52), 8 wk	Clinical assessment, rater	Clinical assessment, rater
Weaver et al.[17]	255	Parkinson disease	Deep brain stimulation v. standard care	UPDRS III (0–108), 6 mo	Examination, neurologist	Examination, neurologist
Noseworthy et al.[18]	168	Multiple sclerosis	Plasma exchange and cyclophosphamide v. placebo	EDSS (0–10), 12 mo	Examination, neurologist	Examination, neurologist
Narins et al.[19]	118	Facial wrinkles	Hyaluronic acid v. collagen	Wrinkle severity rating scale (0–4), 12 wk	Inspection, evaluator	Inspection, treating investigator
Ulm et al.[20]	49	Parkinson disease	Cabergolin v. pergolid	UPDRS III (0–108), 8 wk	Video, investigator	Examination, treating investigator
Meltzer et al.[21,22]	980	Suicide risk	Olanzapine v. Clozapine	CGI-SS, 24 mo	Clinical assessment, psychiatrists	Clinical assessment, psychiatrist
Miller et al.[23]	50	Nasal wound	MeroGel v. Merocel dressing	Synechia (0–3), last follow-up	Endoscopic image, 3 investigators	Live endoscopy, clinician
Taber et al.[24]	26	RSV infection	Ribavirin aerosol v. placebo	Illness severity score (0–3), day 1	Clinical assessment, clinician	Clinical assessment, clinician
US FDA[25]	261	Facial wrinkles	Hylaform v. Zyplast	Severity grading scale (0–5), week 12	Photos, panel	Live inspection, treating investigator
Landsman et al.[26]	36	Onychomycosis	Light therapy v. sham light	CAS* (0–3), day 180	Photos, expert panel	Live inspection, treating physicians
Iglesia et al.[27]	65	Vaginal prolapse	Polypropylene mesh v. standard	POP-Q (0-IV), 3 mo	Examination, practitioner (e.g., nurse, fellow)	Examination, surgeon
Reddihough et al.[28]	61	Cerebral palsy	Botulinum toxin A v. physiotherapy	GMFM (0–264), 6 mo	Video, physiotherapist	Examination,† physiotherapist

Note: CAS = clinical assessment scale, CCSA = Canadian Cardiovascular Society (grading of) angina, CGI-SS = clinical global impression on suicide severity scale (7-point version), EDSS = expanded disability status scale, GMFM = gross motor function measure, PAS = panic and agoraphobia scale, POP-Q = pelvic organ prolapse quantification exam, RSV = respiratory syncytial virus, TMR = transmyocardial laser revascularization, UPDRS = unified Parkinson's disease rating scale, US FDA = US Food and Drug Administration, VAS = visual analogue scale.
*A 4-point grading scale without reported numerical values, which we assigned as 0–3.
†Examinations were performed by 2 assessors for about one-half of the patients (i.e., age < 4 yr).

[CI] –0.40 to –0.06), with moderate heterogeneity ($I^2 = 46\%$, $p = 0.02$) (Figure 3). Thus, the estimated treatment effect based on the assessments of the nonblinded assessors was exaggerated by about one-quarter of the standard deviation of the measurement scale used.

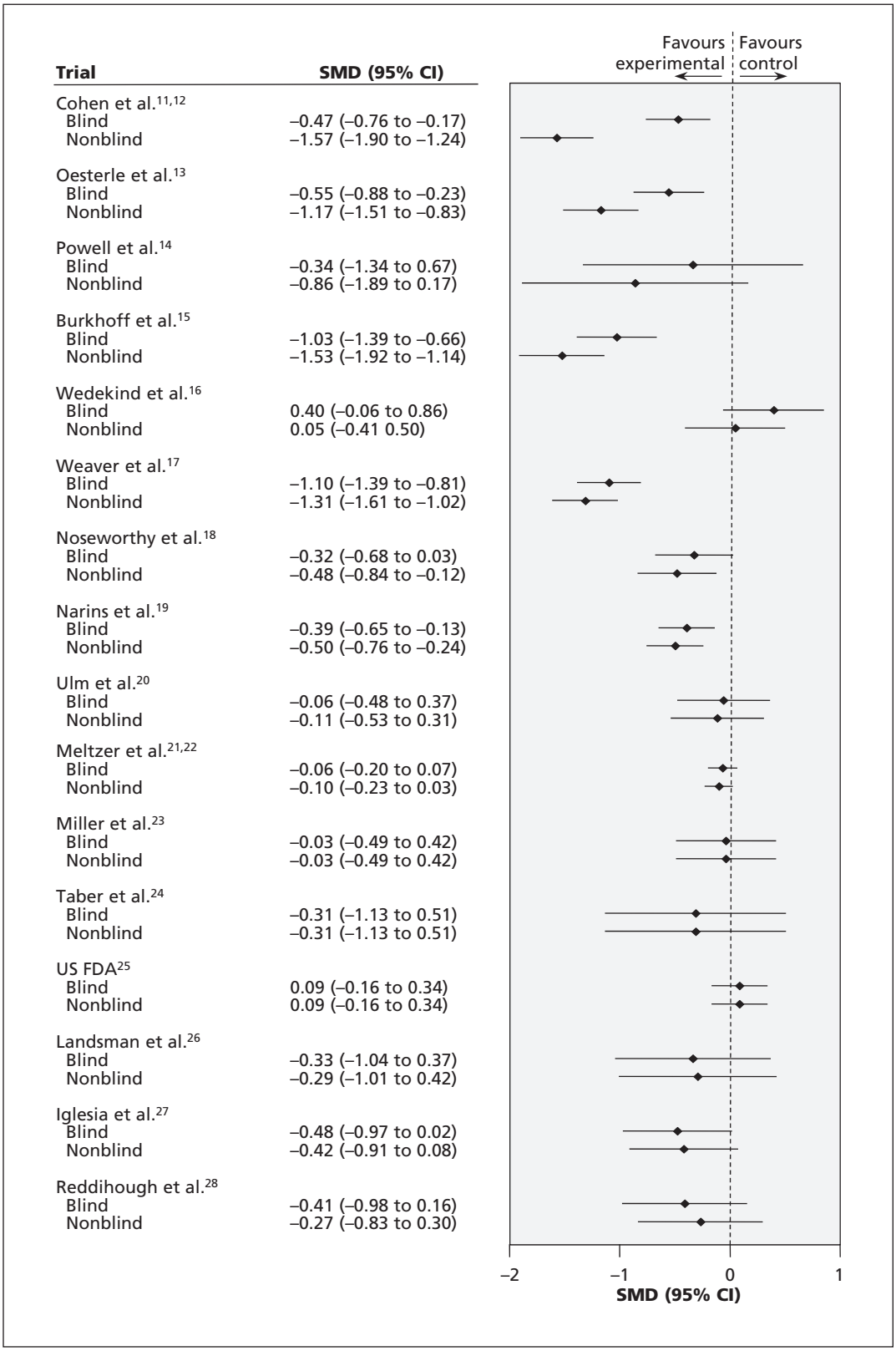

Figure 2: Estimated treatment effect as determined by blinded or nonblinded assessors of outcome. CI = confidence interval, SMD = standard mean difference, US FDA = US Food and Drug Administration.

The pooled effect size based on the assessments of the blinded assessors was –0.34 (95% CI –0.55 to –0.14). Thus, the nonblinded assessors exaggerated the estimated effect size by about 68% (95% CI 14% to 230%) (i.e., –0.23/–0.34 = 0.68).

Our main result was robust, although CIs in our secondary analyses were wide (Table 3). One trial was free from any of the 5 prespecified possible confounders (effect size difference –0.22 [95% CI –0.61 to 0.16].[15] The difference in effect size seemed not to be influenced by any of the suspected confounders (Table 3) or by trial size (data not shown).

Eight trials (involving 980 patients) were included in our review but not in our main meta-analysis because of incomplete or inconsistent data. Qualitative information, or results from other similar trials, suggested notable observer bias in 3 of these trials and no or little bias in 2 trials (Appendix 3).

Using univariable metaregression, we found no statistically significant associations between differences in effect size and high scores for outcome subjectivity (p = 0.29), the degree to which the nonblinded assessors were involved in the trials (p = 0.64), or the vulnerability of the outcome to nonblinded patients (p = 0.80). However, the slope of the regression line between differences in effect sizes and scores for outcome subjectivity was in the expected direction (data not shown). The 13 trials with clearly subjective outcomes had a pooled effect size difference of –0.29 (–0.50 to –0.08) (data not shown). The 3 trials with moderately subjective outcomes had a pooled effect size difference of –0.04 (–0.32 to 0.25) (data not shown).

Interpretation

Nonblinded assessors of subjective measurement scale outcomes in randomized clinical trials tended to generate substantially biased effect sizes. Standardized mean differences were exaggerated by a pooled standard deviation of 0.23 (95% CI 0.40 to 0.06) or, in relative terms, by 68% (95% CI 14% to 230%).

Observer bias can be perceived as the result of the interaction between observers' predispositions and the subjectivity of the outcome. Predis-

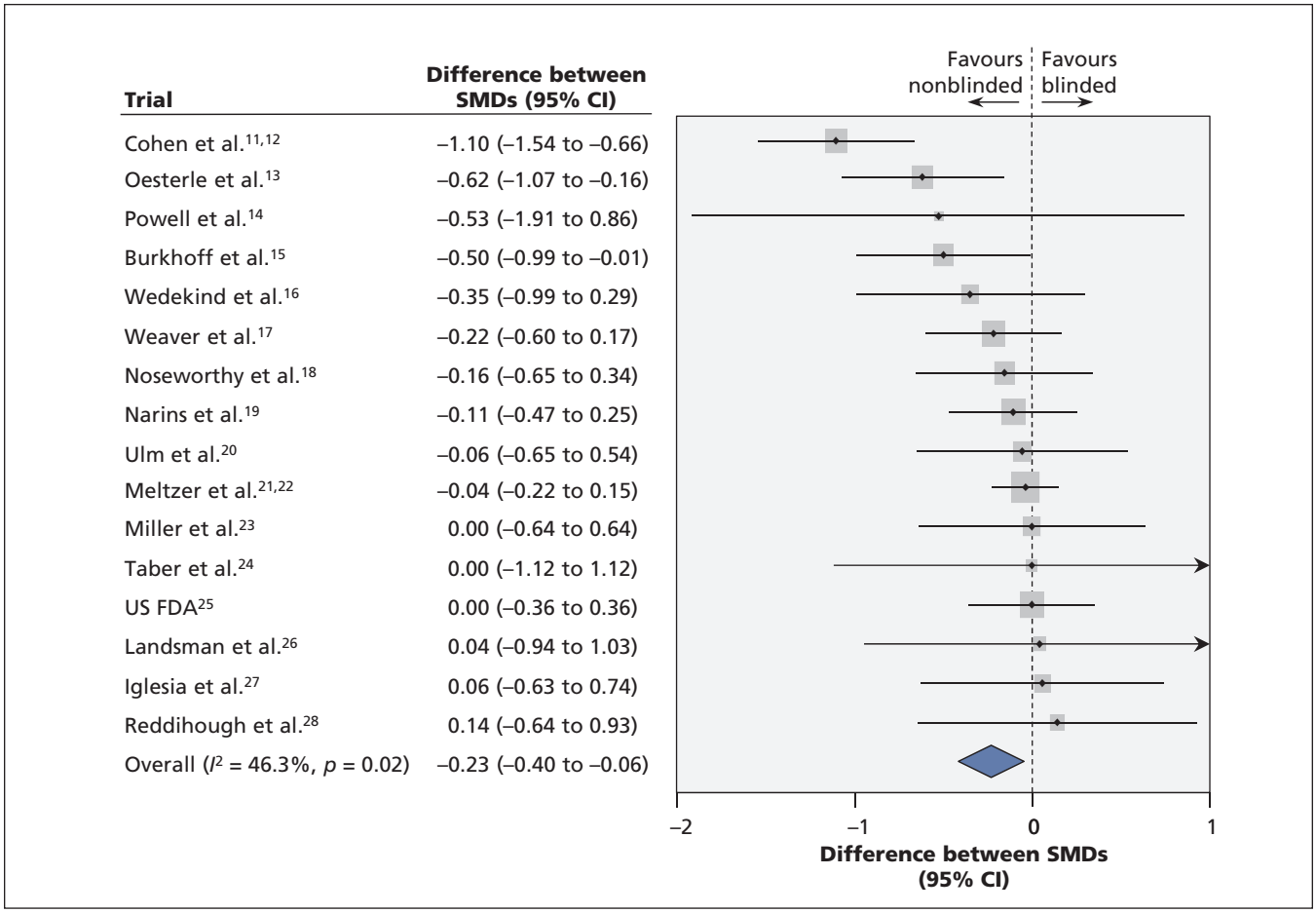

Figure 3: The effect of nonblinded assessors on estimated treatment effects in randomized clinical trials with subjective measurement scale outcomes. Weights were calculated using random effects analysis. CI = confidence interval, SMD = standard mean difference, US FDA = US Food and Drug Administration.

Table 3: Sensitivity and subgroup analyses

Comparison	No. of trials	I^2		Difference* between SMDs (95% CI)
		%	*p* value	
Main analysis	16	46	0.02	–0.23 (–0.40 to –0.06)
Type of analysis				
SMD standardized by SD of blinded control group	16	53	0.007	–0.26 (–0.44 to –0.08)
SMD standardized separately for blinded and nonblinded assessors	16	47	0.02	–0.23 (–0.40 to –0.06)
Correlation accounted for by correlation coefficient	7	88	< 0.001	–0.36 (–0.64 to –0.08)
Correlation accounted for by correlation coefficient or median correlation coefficient†	16	78	< 0.001	–0.21 (–0.35 to –0.07)
Increased precision in crossover/split-body trials was accounted for‡	16	47	0.02	–0.22 (–0.39 to –0.06)
All trials given same weight	16		NA	–0.21 (NA)
Type of data				
Individual patient data	1		NA	–0.16 (–0.65 to 0.34)
Correlation data with no individual patient data	6	60	0.03	–0.46 (–0.83 to –0.10)
Basic outcome data with no information on correlation	9	0	> 0.9	–0.06 (–0.20 to 0.07)
Clinical condition				
Facial wrinkles	3	88	< 0.001	–0.39 (–1.03 to 0.25)
Angina pectoris	2	0	0.7	–0.56 (–0.90 to –0.23)
Parkinson disease	2	0	0.7	–0.17 (–0.49 to 0.15)
Other	9	0	> 0.9	–0.06 (–0.21 to 0.10)
Trial characteristics				
Nonblind assessment by multiple observer consensus	1		NA	0.14 (–0.65 to 0.94)
Nonblind assessment by single observer	15	49	0.02	–0.25 (–0.42 to –0.07)
Publication status: observer bias main objective	4	0	0.8	–0.07 (–0.41 to 0.28)
Publication status: observer bias not main objective	12	58	0.01	–0.27 (–0.47 to –0.06)
Design: parallel group	13	56	0.01	–0.27 (–0.49 to –0.06)
Design: crossover/split-body	3	0	> 0.9	–0.08 (–0.35 to 0.20)
Funding: industry	10	66	0.002	–0.28 (–0.52 to –0.04)
Funding: noncommercial or unclear source	6	0	0.9	–0.13 (–0.38 to 0.12)
Risk of confounding				
Timing of blind and nonblind assessment: same/similar	16	46	0.02	–0.23 (–0.40 to –0.06)
Timing of blind and nonblind assessment: not same/similar	0		NA	NA
Assessors: same type (e.g., neurologists v. neurologists)	10	0	0.6	–0.13 (–0.27 to 0.01)
Assessors: not same (e.g., neurologists v. physiotherapists)	6	73	0.002	–0.30 (–0.69 to 0.09)
Procedure: same type (e.g., clinical exam v. clinical exam)	6	0	0.9	–0.09 (–0.24 to 0.06)
Procedure: not same type (e.g., clinical exam v. video of clinical exam)	10	60	0.008	–0.30 (–0.57 to –0.02)
Blinding procedures: probably effective	11	50	0.03	–0.19 (–0.44 to 0.06)
Blinding procedures: possibly not effective	5	48	0.1	–0.28 (–0.54 to –0.02)
Patients: all seen by both blinded and nonblinded assessors	6	0	0.5	–0.29 (–0.49 to –0.09)
Patients: a minority seen only by one type of assessor	10	59	0.009	–0.20 (–0.46 to 0.06)

Note: CI = confidence interval, NA = not available, SD = standard deviation, SMD = standardized mean difference.
*Pooled difference between SMD based on blind assessments and the corresponding SMD based nonblind assessments. dSMD < 0 suggests than nonblind assessors provide more optimistic estimates of the intervention's effect.
†Seven trials reported a correlation coefficient between blind and nonblind assessments. We used the median correlation coefficient as a correction factor in the 9 trials without such information.
‡Crossover/split-body trials were assigned more weight when their standard error was calculated based on the paired difference (in our planned main analysis, all studies were handled as parallel group trials).

positions are likely to differ substantially from observer to observer and from trial to trial. In some trials, conscientious nonblinded assessors may overcompensate for an expected bias in favour of the experimental intervention and paradoxically induce a bias favouring the control, whereas other trials will have fairly neutral assessors with no important bias. Thus, the degree of observer bias in trials with clearly predisposed outcome assessors is likely to be considerably higher than the mean we see here, which is based on all of the included trials. When determining the risk of bias attributable to nonblinded assessors in a randomized trial, we suggest being mindful of the range of observer bias we have found, and not only the pooled mean.

Based largely on convention, standardized mean differences of –0.2 are considered small effects, –0.5 are considered medium effects, and –0.8 are considered large effects.[37] By such standards, our result constitutes a small to moderate difference. However, it seems inappropriate to interpret a degree of bias in the same way as we would interpret a treatment effect. The relevant problem is how much bias can be expected when using a nonblinded assessor, not whether that degree of bias represents a clinically worthwhile effect. In a situation with a large true treatment effect with a standardized mean difference of –0.8, the average degree of observer bias when using nonblinded observers, –0.23, would imply an exaggeration of the treatment effect estimate by 29%. This percentage increases to 115% if effects are small (i.e., if the standardized mean difference is –0.2). In the 16 trials we analyzed, the pooled estimated treatment effect was exaggerated by 68% (14% to 230%) when based on data from nonblinded assessors. Thus, we interpret our result as evidence for a substantial degree of observer bias.

In a Cochrane review of the effect of progressive resistance strength training, Liu and colleagues compared pooled standardized mean differences in a subgroup of 54 randomized trials using nonblinded assessors (–0.88 [95% CI –0.77 to –0.99]) with that of 19 trials using blinded assessors (–0.23 [95% CI –0.13 to –0.34]).[4,38] The result of this indirect comparison is within the range of our findings. Meta-epidemiological studies of trials with binary outcomes have reported inconsistent estimates of the effect of a lack of double-blinding.[5] However, our result is consistent with that of Savovic and colleagues,[6] and with our previous study of observer bias in trials with binary outcomes.[8]

It may be tempting to use measures for interobserver agreement (e.g., weighted κ, intraclass correlation coefficients) as surrogate markers for risk of observer bias. Similarly, training nonblinded observers to reduce interobserver variation[39] could be seen as an appealing alternative to blinding in a situation where blinding is challenging. However, good interobserver agreement does not prevent observer bias. For example, the trial with the largest degree of observer bias[11] used a scale reported to have an intraclass correlation coefficient as high as 0.87.[40]

Some researchers consider the blinding of outcome assessors too resource-demanding, superfluous, or misconceived;[41,42] however, planning and running a randomized clinical trial is already a logistically very challenging undertaking. The comparatively minor investment of using blinded outcome assessors reduces the risk of bias considerably. Blinding outcome assessors is possible in most trials.[43,44]

Limitations

The trials we included in our analysis are contemporary and represent a variety of clinical specialties, and their design implies a low risk of confounding. However, these trials are not representative of medical trials in general. We included no trials with clearly objective measurement scale outcomes, such as nonrepeatable automatized laboratory measures. The included trials had subjective outcomes, and our results apply only to similar trials. Furthermore, extrapolating our results to all trials with subjective measurement scale outcomes assumes that trials with both blinded and nonblinded assessors are comparable with trials with only nonblinded assessors.

Our preplanned main analysis disregarded the correlation between blinded and nonblinded assessments, and its confidence interval may thus be somewhat inflated. However, the correlation was available for 7 trials, and secondary analyses incorporating the correlation between blinded and nonblinded assessments provided results similar to those of the main analysis.

Because searching for trials with both blinded and nonblinded assessors is challenging, some such studies may not have been identified by our literature search. However, it is unclear whether such trials would report substantially different results. Publication bias is normally driven by the effect of a treatment[45] and may have a limited, yet unpredictable, effect on our comparison between types of assessments.

Conclusion

We provide empirical evidence for observer bias in randomized clinical trials with subjective measurement scale outcomes. Failure to blind outcome assessors in such trials results in a high risk of substantial bias.

References

1. Higgins JPT, Green S, editors. *Cochrane handbook for systematic reviews of interventions*. Oxford (UK): The Cochrane Collaboration; 2011.
2. Haahr MT, Hróbjartsson A. Who is blind in randomised clinical trials? An analysis of 200 trials and a survey of authors. *Clin Trials* 2006;3:360-5.
3. Poolman RW, Struijs PA, Krips R, et al. Reporting of outcomes in orthopaedic randomized trials: Does blinding of outcome assessors matter? *J Bone Joint Surg Am* 2007;89:550-8.
4. Liu CJ, LaValley M, Latham NK. Do unblinded assessors bias muscle strength outcomes in randomized controlled trials of progressive resistance strength training in older adults? *Am J Phys Med Rehabil* 2011;90:190-6.
5. Pildal J, Hróbjartsson A, Jørgensen KJ, et al. Impact of allocation concealment on conclusions drawn from meta-analyses of randomised trials. *Int J Epidemiol* 2007;36:847-57.
6. Savovic J, Jones HE, Altman DG, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. *Ann Intern Med* 2012;157:429-38.
7. Devereaux PJ, Manns BJ, Ghali WA, et al. Physician interpretations and textbook definitions of blinding terminology in randomized controlled trials. *JAMA* 2001;285:2000-3.
8. Hróbjartsson A, Thomsen AS, Emanuelsson F, et al. Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. *BMJ* 2012;344:e1119.
9. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7:177-88.
10. Fieller EC. Some problems in interval estimation. *J R Stat Soc [Ser A]* 1954;16:175-85.
11. Cohen SR, Holmes RE. Artecoll: a long-lasting injectable wrinkle filler material: report of a controlled, randomized, multicenter clinical trial of 251 subjects. *Plast Reconstr Surg* 2004;114:964-76, discussion 977-9.
12. Cohen SR. FDA summary of safety and effectiveness data. Silver Spring (MD): US Food and Drug Administration; 2004. Available: www.accessdata.fda.gov/cdrh_docs/pdf2/P020012b.pdf (accessed 2011 Aug. 5).
13. Oesterle SN, Sanborn TA, Ali N, et al. Percutaneous transmyocardial laser revascularisation for severe angina: the PACIFIC randomised trial. Potential class improvement from intramyocardial channels. *Lancet* 2000;356:1705-10.
14. Powell NB, Zonato AI, Weaver EM, et al. Radiofrequency treatment of turbinate hypertrophy in subjects using continuous positive airway pressure: a randomized, double-blind, placebo-controlled clinical pilot trial. *Laryngoscope* 2001;111:1783-90.
15. Burkhoff D, Schmidt S, Schulman SP, et al. Transmyocardial laser revascularisation compared with continued medical therapy for treatment of refractory angina pectoris: a prospective randomised trial. ATLANTIC Investigators. Angina treatments — lasers and normal therapies in comparison. *Lancet* 1999;354:885-90.
16. Wedekind D, Broocks A, Weiss N, et al. A randomized, controlled trial of aerobic exercise in combination with paroxetine in the treatment of panic disorder. *World J Biol Psychiatry* 2010;11:904-13.
17. Weaver FM, Follett K, Stern M, et al.; CSP468 Study Group. Bilateral deep brain stimulation vs best medical therapy for patients with advanced Parkinson disease: a randomized controlled trial. *JAMA* 2009;301:63-73.
18. Noseworthy JH, Vandervoort MK, Penman M, et al. Cyclophosphamide and plasma exchange in multiple sclerosis. *Lancet* 1991;337:1540-1.
19. Narins RS, Coleman W, Donofrio L, et al. Nonanimal sourced hyaluronic acid–based dermal filler using a cohesive polydensified matrix technology is superior to bovine collagen in the correction of moderate to severe nasolabial folds: results from a 6-month, randomized, blinded, controlled, multicenter study. *Dermatol Surg* 2010;36:730-40.
20. Ulm G, Schüler P. Cabergolin versus pergolid: a video-blinded, randomised multicenter cross-over study. *Akt Neurol* 1999;26:360-5.
21. Meltzer HY, Alphs L, Green AI, et al. Clozapine treatment for suicidality in schizophrenia: International Suicide Prevention Trial (InterSePT). *Arch Gen Psychiatry* 2003;60:82-91.
22. Meltzer HY. FDA statistical review and evaluation. Silver Spring (MD): US Food and Drug Administration; 2003. Available: www.fda.gov/ohrms/dockets/ac/02/briefing/3908B1_02_E-%20Statistical%20Review.pdf (accessed 2013 Jan. 8).
23. Miller RS, Steward DL, Tami TA, et al. The clinical effects of hyaluronic acid ester nasal dressing (Merogel) on intranasal wound healing after functional endoscopic sinus surgery. *Otolaryngol Head Neck Surg* 2003;128:862-9.
24. Taber LH, Knight V, Gilbert BE, et al. Ribavirin aerosol treatment of bronchiolitis associated with respiratory syncytial virus infection in infants. *Pediatrics* 1983;72:613-8.
25. US Food and Drug Administration. FDA summary of safety and effectiveness data: Hylaform. Silver Spring (MD): The Administration; 2004. Available: www.accessdata.fda.gov/cdrh_docs/pdf3/P030032b.pdf (accessed 2011 Aug. 5).
26. Landsman AS, Robbins AH, Angelini PF, et al. Treatment of mild, moderate, and severe onychomycosis using 870- and 930-nm light exposure. *J Am Podiatr Med Assoc* 2010;100:166-77.
27. Iglesia CB, Sokol AI, Sokol ER, et al. Vaginal mesh for prolapse: a randomized controlled trial. *Obstet Gynecol* 2010;116:293-303.
28. Reddihough DS, King JA, Coleman GJ, et al. Functional outcome of botulinum toxin A injections to the lower limbs in cerebral palsy. *Dev Med Child Neurol* 2002;44:820-7.
29. Baumann LS, Shamban AT, Lupo MP, et al. JUVEDERM vs. ZYPLAST Nasolabial Fold Study Group. Comparison of smooth-gelhyaluronic acid dermal fillers with cross-linked bovine collagen: a multicenter, double-masked, randomized, within-subject study. *Dermatol Surg* 2007;33(Suppl 2):S128-35.
30. Purdue GF, Hunt JL, Still JM Jr, et al. A multicenter clinical trial of a biosynthetic skin replacement, Dermagraft-TC, compared with cryopreserved human cadaver skin for temporary coverage of excised burn wounds. *J Burn Care Rehabil* 1997;18:52-7.
31. Herberger K, Franzke N, Blome C, et al. Efficacy, tolerability and patient benefit of ultrasound-assisted wound treatment versus surgical debridement: a randomized clinical study. *Dermatology* 2011;222:244-9.
32. Alam M, Pon K, Van Laborde S, et al. Clinical effect of a single pulsed dye laser treatment of fresh surgical scars randomized controlled trial. *Dermatol Surg* 2006;32:21-5.
33. Ash K, Lord J, Zukowski M, et al. Comparison of topical therapy for striae alba (20% glycolic acid/0.05% tretinoin versus 20% glycolic acid/10% L-ascorbic acid). *Dermatol Surg* 1998;24:849-56.
34. Realmuto GM, Erickson WD, Yellin AM, et al. Clinical comparison of thiothixene and thioridazine in schizophrenic adolescents. *Am J Psychiatry* 1984;141:440-2.
35. Havel CJ Jr, Strait RT, Hennes H. A clinical trial of propofol vs. midazolam for procedural sedation in a pediatric emergency department. *Acad Emerg Med* 1999;6:989-97.
36. Kadish A, Nademanee K, Volosin K, et al. A randomized controlled trial evaluating the safety and efficacy of cardiac contractility modulation in advanced heart failure. *Am Heart J* 2011;161:329-337.e1-2.
37. Cohen J. *Statistical power analysis for the behavioral sciences*. New York (NY): Academic Press; 1977.
38. Liu CJ, Latham NK. Progressive resistance strength training for improving physical function in older adults. *Cochrane Database Syst Rev* 2009;(3):CD002759.
39. Brorson S, Bagger J, Sylvest A, et al. Improved interobserver variation after training of doctors in the Neer system. A randomised trial [published erratum in *J Bone Joint Surg Br* 2003;85:153]. *J Bone Joint Surg Br* 2002;84:950-4.
40. Lemperle G, Holmes RE, Cohen SR, Lemperle SM. A classification of facial wrinkles. *Plast Reconstr Surg* 2001;108:1735-50; discussion 1751-2.
41. Dodd DC. Blind slide reading or the uninformed versus the informed pathologist. *Comments Toxicol* 1988;2:88-91.
42. Burkhardt JE, Ennulat D, Pandher K, et al. Topic of histopathology blinding in nonclinical safety biomarker qualification studies. *Toxicol Pathol* 2010;38:666-7.
43. Boutron I, Guittet L, Estellat C, et al. Reporting methods of blinding in randomized controlled trials assessing non-pharmacological treatments. A systematic review. *PLoS Med* 2007;4:e61.
44. Karanicolas PJ, Bhandari M, Walter SD, et al.; Collaboration for Outcomes Assessment in Surgical Trials (COAST) Musculoskeletal Group. Radiographs of hip fractures were digitally altered to mask surgeons to the type of implant without compromising the reliability of quality ratings or making the rating process more difficult. *J Clin Epidemiol* 2009;62:214-223.e1.
45. Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ* 1997;315:640-5.

Affiliations: From the Nordic Cochrane Centre (Hróbjartsson, Thomsen, Emanuelsson, Tendal) Rigshospitalet Department 7811, Copenhagen, Denmark; the Department of Biostatistics (Hilden), University of Copenhagen, Denmark; the

French Cochrane Centre (Boutron, Ravaud), Assistance Publique (Hotel Dieu), Paris, France; and the Department of Orthopaedic Surgery (Brorson), Herlev University Hospital, Denmark.

Contributors: Asbjørn Hróbjartsson conceived the idea and design of the study, organized the study and wrote the first draft of the manuscript. Ann Thomsen and Asbjørn Hróbjartsson developed the search strategy. Ann Thomsen, Frida Emanuelsson, Britta Tendal, Stig Brorson and Asbjørn Hróbjartsson did the nonmasked data collection. Isabelle Boutron, Philippe Ravaud, Stig Brorson, Britta Tendal and Asbjørn Hróbjartsson did the masked data collection. Asbjørn Hróbjartsson and Jørgen Hilden did the statistical analyses. All of the authors revised the manuscript for important intellectual content and approved the final version submitted for publication.

Funding: The study was partially funded by the Danish Council for Independent Research: Medical Sciences. The funder had no influence on the study's design, the collection, analysis, and interpretation of data, or the writing of the article and the decision to submit it for publication.

Acknowledgements: The authors thank the following trial authors for sharing unpublished outcome data: Peggy Vandervoort, George C. Ebers, Daniel Burkhoff, Cheryl Iglesia, Borwin Bandelow and Dina S. Reddihough, and Frances S. Weaver and the US Department of Veterans Affairs (VA) Cooperative Study Program, as well as the VA CSP study #468 "A comparison of best medical therapy and deep brain stimulation of subthalamic nucleus and globus pallidus for the treatment of Parkinson's disease." The authors also thank Peter C. Gøtzsche and Andreas Lundh for valuable comments on previous versions of the manuscript.