International Journal of Epidemiology, 2014, 937–948
doi: 10.1093/ije/dyt270
Advance Access Publication Date: 21 January 2014
Original article

Original article

Observer bias in randomized clinical trials with time-to-event outcomes: systematic review of trials with both blinded and non-blinded outcome assessors

Asbjørn Hróbjartsson,[1]*, Ann Sofia Skou Thomsen,[1] Frida Emanuelsson,[1] Britta Tendal,[1] Jeppe Vejlgaard Rasmussen, [2] Jørgen Hilden,[3] Isabelle Boutron,[4] Philippe Ravaud[4] and Stig Brorson[2]

[1]Nordic Cochrane Centre, Rigshospitalet Department 7811, Copenhagen, Denmark, [2]Department of Orthopaedic Surgery, Herlev University Hospital, Copenhagen, Denmark, [3]Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark and [4]French Cochrane Centre, Assistance Publique (Hotel Dieu), Université Paris Descartes, Paris, France

*Corresponding author. Nordic Cochrane Centre, Rigshospitalet Department 7811, Blegdamsvej 9, 2100 Copenhagen Ø, Denmark. E-mail: ah@cochrane.dk

Accepted 9 December 2013

Abstract

Background: We wanted to evaluate the impact of nonblinded outcome assessors on estimated treatment effects in time-to-event trials.

Methods: Systematic review of randomized clinical trials with both blinded and non-blinded assessors of the same time-to-event outcome. Two authors agreed on inclusion of trials and outcomes. We compared hazard ratios based on nonblinded and blinded assessments. A ratio of hazard ratios (RHR) <1 indicated that nonblinded assessors generated more optimistic effect estimates. We pooled RHRs with inverse variance random-effects meta-analysis.

Results: We included 18 trials. Eleven trials (1969 patients) with subjective outcomes provided hazard ratios, RHR 0.88 (0.69 to 1.12), ($I^2 = 44\%$, $P = 0.06$), but unconditional pooling was problematic because of qualitative heterogeneity. Four atypical cytomegalovirus retinitis trials compared experimental oral administration with control intravenous administration of the same drug, resulting in bias favouring the control intervention, RHR 1.33 (0.98 to 1.82). Seven trials of cytomegalovirus retinitis, tibial fracture and multiple sclerosis compared experimental interventions with standard control interventions, e.g. placebo, no-treatment or active control, resulting in bias favouring the experimental intervention, RHR 0.73 (0.57 to 0.93), indicating an average exaggeration of nonblinded hazard ratios by 27% (7% to 43%).

Conclusions: Lack of blinded outcome assessors in randomized trials with subjective time-to-event outcomes causes high risk of observer bias. Nonblinded outcome

assessors typically favour the experimental intervention, exaggerating the hazard ratio by an average of approximately 27%; but in special situations, nonblinded outcome assessors favour control interventions, inducing a comparable degree of observer bias in the reversed direction.

Key words: Randomized clinical trials, time-to-event, blinding, bias, observer bias

Key Messages

- Nonblinded outcome assessors are sometimes used in randomized time-to-event trials with subjective outcomes, but the typical degree of observer bias is not known.
- We identified and reviewed time-to-event trials with both blinded and nonblinded outcome assessment, as such trials provide a reliable evaluation of observer bias.
- In 18 trials, of which 11 were included in the main meta-analysis, nonblinded observers tended to favour the experimental intervention, and exaggerated the hazard ratio by approximately 27%.
- Under special circumstances nonblinded observers may favour the control intervention and will tend to underestimate the hazard ratio by a similar extent.
- Generalizability of the findingis hampered by the moderate numbers of trials and outcomes.

Introduction

Randomized clinical trials with time-to-event outcomes are commonly conducted without blinded outcome assessors.[1–4] In trials with objective outcomes, for example time-to-death, bias seems unlikely. In trials with subjective outcomes, i.e. involving assessor judgement,[5] it is prudent to suspect observer bias.

Observer bias is generated by the conscious or unconscious predispositions of nonblinded observers, for example due to hope or expectations. When the trial outcome is based on the classification of events, as in time-to-event outcomes, the bias is mediated by directional misclassification, i.e. the blinded and nonblinded assessments differ systematically and not only randomly, and will often favour the experimental intervention.

Use of blinded outcome assessors protects against observer bias but will usually increase cost and logistical complexity of a trial with no guarantee that blinding will make an important difference in any single trial. Some commentators find blinded outcome assessment superfluous or wrong.[6–8] For example, the Pharmaceutical Research and Manufacturers of America's working group discouraged the standard use of central blinded review of progression-free survival in cancer trials.[8]

Unfortunately, the empirical basis for blinding outcome assessors in subjective time-to-event trials is surprisingly incomplete. Studies of blinding include case histories which may not be representative[4] and indirect comparisons of double-blinded trials with similar trials not reported as 'double-blinded'.[9,10] Such meta-epidemiological studies have a notable risk of confounding, because trials that are reported as double-blinded may differ from trials that are not for many other reasons, for example sample size, concealment of the allocation sequence, funding source or multi-centre status. This problem is augmented by the limited number of adjustments possible for suspected confounders. Meta-epidemiological studies are also challenging to interpret because double-blinded is an ambiguous term.[1,11] Finally, meta-epidemiological studies typically address binary outcomes, not time-to-event outcomes.

Empirical analyses of observer bias (also called 'detection bias' or 'ascertainment bias') serve as a guide to the design of future trials and to the assessment of risk of bias in conducted trials, for example when a trial is included in a meta-analysis.[5] The most reliable design for studies of observer bias involves trials that use both blinded and nonblinded assessors of the same outcome because the direct comparison minimizes risk of confounding. In two analyses of such studies we found evidence of observer bias in trials with binary outcomes[12] and measurement scale outcomes.[13]

In the present systematic review we investigate observer bias in time-to-event trials. Our primary objective was to evaluate the impact of nonblinded outcome assessment on estimated treatment effects in randomized clinical

time-to-event trials; our secondary objective was to examine reasons for variation in degree of observer bias.

Methods

We included randomized clinical trials with blinded and nonblinded assessment of the same time-to-event outcome. A time-to-event outcome was defined as involving the evaluation of time from randomization to clinical event based on five time points or more.

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 expected direction of any bias (e.g. a trial that compared two treatments in common use without designating one as 'experimental' and the other as 'control'). We also excluded: trials where only a subgroup of patients had been evaluated by blinded and nonblinded assessors, unless the subgroup was selected at random; trials where blinded and nonblinded assessors had access to each other's results; and trials where initially blinded assessors clearly had become unblinded, e.g. when radiographs showed ceramic material indicative of the experimental intervention. Finally, we excluded trials with blinded end-point committees adjudicating the assessments made by nonblinded clinicians, because such adjudication often involves foreknowledge of the nonblinded assessment or is restricted to adjudication of events only.

We searched: PubMed, EMBASE, PsycINFO, CINAHL, the Cochrane Central Register of Controlled Trials, HighWire Press and Google Scholar (Appendix 1, available as Supplementary data at *IJE* online). The last formal search was performed on 15 September 2013. References of all included trials were read.

One author read all abstracts from standard databases and all text fragments from full-text databases. If a study might be eligible, a full study report was retrieved and read by one author who excluded all clearly ineligible studies. Two authors read all remaining study reports and decided on eligibility. Disagreements were resolved by discussion.

We extracted background data, and outcome data resulting from both the blinded assessment and the nonblinded assessment. If data were incomplete, we e-mailed and/or telephoned the corresponding authors. We also searched the US Food and Drug Administration (FDA) website. For trials with more than two groups, we pooled the results in the experimental or the control groups, when possible. When authors chose to send us individual patient data, we checked whether all randomized patients were included in the dataset and tried to replicate a result of the published paper. Two authors independently derived outcome data. Any discrepancy was solved by discussion.

For each trial, we evaluated four pre-specified factors that could systematically distort the measured differences between blinded and nonblinded assessors: (i) a considerable time difference between the blinded and nonblinded assessments; (ii) different types of assessors (e.g. nurses vs physicians); (iii) different types of procedures (e.g. direct visual assessment of wound vs assessment of photographs of wound); d) a substantial risk of ineffective blinding procedure. These measurement distortion factors were evaluated by two masked authors unaware of the result of the trial, as previously reported.[12,13]

Using the same masking procedure, we also evaluated characteristics of each outcome assessment. Two authors scored independently three factors on a 1–5 scale (1 was low and 5 high): (i) the degree of outcome subjectivity (i.e. the degree of assessor judgment); (ii) the nonblinded outcome 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 (iii) the outcome vulnerability to nonblinded patients (e.g. high in outcomes based on interviews with nonblinded patients). Disagreements were resolved by discussion.

Our primary outcome was ratio of hazard ratios (RHRs). We noted, or calculated, the hazard ratio (HR) and its standard error for each trial based on blinded assessors, $HR_{blinded}$, and based on nonblinded assessors, $HR_{nonblinded}$. A HR < 1 indicates a beneficial effect of the experimental intervention in trials assessing time to a harmful event, for example progression of multiple sclerosis. If HR or the standard error was not reported explicitly, we approximated a HR and a standard error based on other available information, for example Kaplan-Meier curves[14] (Appendix 3, available as Supplementary data at *IJE* online). In trials with individual patient data we calculated the HR and the standard error directly based on Cox regression models. In trials assessing time to a harmful event we summarized the impact of nonblinded outcome assessment as the ratio of the hazard ratios, $RHR = HR_{nonblinded} / HR_{blinded}$. For trials assessing time to beneficial events we reversed the ratio, so that for all trials a RHR < 1 indicated that the nonblinded assessors generate more optimistic effect estimates than the blinded assessment. The standard error (SE) of RHR for our main analyses disregarded the dependence between the blinded and nonblinded assessment (Appendix 3, available as Supplementary data at *IJE* online).

We then meta-analysed the individual RHRs with inverse variance methods using random-effects models.[15] The same approach was used for meta-analysing ratio of medians ratio (Appendix 3, available as Supplementary data at *IJE* online). All hazard results rested on log-scale analysis. The statistical software used was Stata 11 and SAS 9.2.

We pooled all trials, and explored whether the effect differed in subgroups of trials divided by: type of data; clinical conditions; consensus status of the nonblinded assessors; the objective of a trial; source of funding; and the risk of measurement distortion. We also used standard errors of the ratio of hazard ratios that incorporated the dependency between blinded and nonblinded assessments in trials with individual patient data (Appendix 3, available as Supplementary data at *IJE* online). For the other trials we used as a correction factor the median reduction of SE in the trials with individual patient data, [(SE ignoring dependence − SE incorporating dependence) / SE ignoring dependence]. Finally, we also assumed equal weights for each trial.

We had planned to explore whether the variation in the ratio of hazard ratios was associated with the three prespecified outcome characteristics described above; however, the scores for the outcome characteristics varied too little for a meaningful analysis (Appendix 4, available as Supplementary data at *IJE* online).

Results

We examined 555 publications based on 3149 hits in standard databases and 3541 hits in full-text databases. We excluded 537 studies, mostly because they were not randomized clinical trials or lacked blinded or nonblinded outcome assessment (Appendix 2, available as Supplementary data at *IJE* online). Thus, 18 trials were included[16–33]) (Table 1).

We had access to both blinded and nonblinded outcome data in 12 trials (2250 patients)[16–27] of which three provided individual patient data;[16–18] eleven trials (1969 patients) provided data for estimation of the ratio of hazard ratios, and also 11 trials (2222 patients) for the ratio of medians ratio. The outcomes were subjective (Table 2); on a scale from 1 to 5 (where 5 indicates high subjectivity) all 12 trials with outcome data scored 4–5. In addition, we had access to qualitative or incomplete data on the blinded and nonblinded outcome assessments in four additional trials.[28–33]

Main outcome: ratio of hazard ratios

In 7 of 11 trials (64%) the hazard ratio was more favourable to the experimental intervention when based on the nonblinded outcome assessors (Figure 1). The ratio of the hazard ratios spanned from 0.25 to 1.56 (Figure 2).

The pooled ratio of hazard ratios for all eleven trials[16–26] was 0.88 (0.69 to 1.12), ($I^2 = 44\%$, $P = 0.06$). The unconditional pooling of all trials was not meaningful, however, due to qualitative heterogeneity driven by four of the seven cytomegalovirus retinitis trials ($I^2 = 63\%$, $P = 0.01$) (Figure 2).

Table 1. Characteristics of 18 randomized clinical trials included

Characteristic	Number
General	
Parallel group design	17
Number of study groups: 2	9
Primary outcome defined	17
Intervention: drug	14
Intervention: surgery/procedure	4
Control group: usual care/active control	15
Control group: placebo/no-treatment	3
[a]Published by specialty journal (e.g. *Arch Ophthalmol*)	10
[a]Published by general medical journal (e.g. *Lancet*)	7
Outcome	
Clearly subjective (score 4-5 on a 1-5 scale)	17
[b]Moderately subjective (score 2-3)	1
Objective (score 1)	0
Medical specialty	
Ophthalmology	9
Orthopaedic surgery	4
Surgery	2
Neurology, oncology, dermatology	3
Trial methodology	
Random allocation sequence adequately generated	6
Random allocation sequence adequately concealed	8
Patients blind	3
Treatment provider blind	0

[a]One trial was unpublished.
[b]The single trial with an outcome that was moderately subjective had incompletely reported outcomes and was not included in our meta-analyses.

Seven standard trials included three cytomegalovirus retinitis trials,[20–22] three trials of tibial fractures[17–19] and one trial of multiple sclerosis.[16] All compared experimental interventions with standard control interventions, such as placebo, no-treatment, usual care or active control. The pooled ratio of hazard ratios was 0.73 (0.57 to 0.93), ($I^2 = 24\%$, $P = 0.25$), indicating that nonblinded hazard ratios were exaggerated by 27%, on average (Figure 3).

Four trials were atypical as they compared an oral experimental administration of a drug with the intravenous control administration of the same drug for cytomegalovirus retinitis.[22–25] The pooled ratio of hazard ratios in these trials was 1.33 (0.98 to 1.82), with no heterogeneity ($I^2 = 0\%$, $P = 0.89$), indicating that, on average, nonblinded outcome assessors favoured the control intervention and underestimated the hazard ratio by 33% (-2% to 82%) (Figure 3).

Supplementary outcome: ratio of medians ratio

The same pattern of result was found when we analysed ratio of medians ratios (Appendix 3, available as Supplementary data at *IJE* online).

Table 2. Characteristics of the outcome assessments

Trial	[a]N	Clinical condition	Experimental vs control	Time to event outcome	Assessment	
					Blind	Nonblind
Noseworthy 1991[16]	168	Multiple sclerosis	Plasma exchange & [b]CPM vs placebo	Progression of [c]MS; assessment every 6 months for up to 3 years	Examination, neurologist	Examination, neurologist
Aro 2011[17]	277	Fracture	[d]rhBMP-2 vs usual care only	Healed fracture; assessments weeks 6, 10, 13, 16, 20, 24, 32, 41	Radiograph, radiologist	Radiograph, surgeon
Govender 2002[18]	450	Fracture	rhBMP-2 vs usual care only	Healed fracture; assessments weeks 6, 10, 14, 20, 26, 39, 50	Radiograph, radiologist	Radiograph, surgeon
Liebergall 2013[19]	24	Fracture	Stem cell vs usual care only	Healed fracture; assessments weeks 6, 13, 26, 39, 52	Radiograph, surgeon	Radiograph, surgeon
HPMPC 1997[20]	64	[e]CMV retinitis	Cidofovir: [f]iv vs no-treatment	Progression of retinitis, assessment bi-weekly to week 11, then every 4 weeks	Retinal photos, graders	Retinal photos, ophthalmologists
CRT 1996 [21]	279	CMV retinitis	[g]Combined & Ganci vs Foscarnet	Progression of retinitis, assessment monthly x 6, then every 3 months	Retinal photos, graders	Retinal photos, ophthalmologists
Vitravene 2002[22]	28	CMV retinitis	Fomiversen iv vs no-treatment	Progression of retinitis, assessment [h]weekly (control) for up to 18 weeks	Retinal photos, graders	Retinal photos, ophthalmologists
Drew1995[23]	123	CMV retinitis	Oral vs iv ganciclovir	Progression of retinitis, assessment every 2 weeks for up to 20 weeks	Retinal photos, single grader	Fundoscopy, ophthalmologists
Danner 1995[24]	159	CMV retinitis	Oral vs iv ganciclovir	Progression of retinitis, assessment every 2 weeks for up to 20 weeks	Retinal photos, single grader	Fundoscopy, ophthalmologists
Squires B 1996[25]	237	CMV retinitis	Oral vs iv ganciclovir	Progression of retinitis, assessment every 2 weeks for up to 20 weeks	Retinal photos, single grader	Fundoscopy, ophthalmologists
Martin 2002[26]	160	CMV retinitis	Oral valganciclovir vs iv gangciclovir	Progression of retinitis, assessment every two weeks through week 16	Retinal photos, a grader	Fundoscopy, ophthalmologists
Lalezari 2002[27]	281	CMV retinitis	Oral dose escalation vs iv ganciclovir	Progression of retinitis, assessment every 2 weeks for up to 26 weeks	Retinal photos, graders	Fundoscopy, ophthalmologists
[i]Still 2003[28]	82	Wound	Collagen sponge vs usual dressing	Healed wound; assessment days 3, 7 and every 2 days until day 32	Photo, 3 burn experts	Examination, clinician
[i]Swiontkowski 2006[29]	60	Fracture	rhBMP-2 vs usual care only	Healed fracture; assessments weeks 6, 10, 14, 20, 26, 39, 50	Radiograph, radiologist	Radiograph, surgeon
[i]Novo TTF[30]	237	[j]GBM	Novo TTF vs usual care	Radiological progression; assessments every 2 months for up to 40 months	[k]MRI, radiologist	MRI, clinician
[i]Dumville 2009[31]	267	Venous ulcer	Larvae vs wound dressing	Healed ulcer, assessments weekly for 1/2 year, then monthly for up to 1 year	Photo, 2 assessors	Live inspection, 2 nurses
[l]Musch 1997[32]	188	CMV retinitis	Ganciclovir: implant vs iv	Progression of retinitis, assessment bi-weekly through week 8, then monthly	Retinal photos, graders	Fundoscopy, ophthalmologists
[l]Kabat-Zinn 1998[33]	37	Psoriasis	Stress reduction vs usual care	Clearing of psoriasis lesion, assessment every 4 weeks	Live inspection, physicians	Live inspection, nurse

[a]N, number of patients randomized; [b]cyclophosphamide; [c]multiple sclerosis; [d]recombinant human bone morphogenetic protein-2; [e]cytomegalovirus; [f]intravenous or intravitreous; [g]ganciclovir + foscarnet vs ganciclovir vs foscarnet; [h]every 2 weeks for the experimental group; [i]incomplete or inconsistent outcomes; [j]glioblastoma multiforme; [k]magnetic resonance imaging; [l]no data to provide information on size or direction of observer bias.

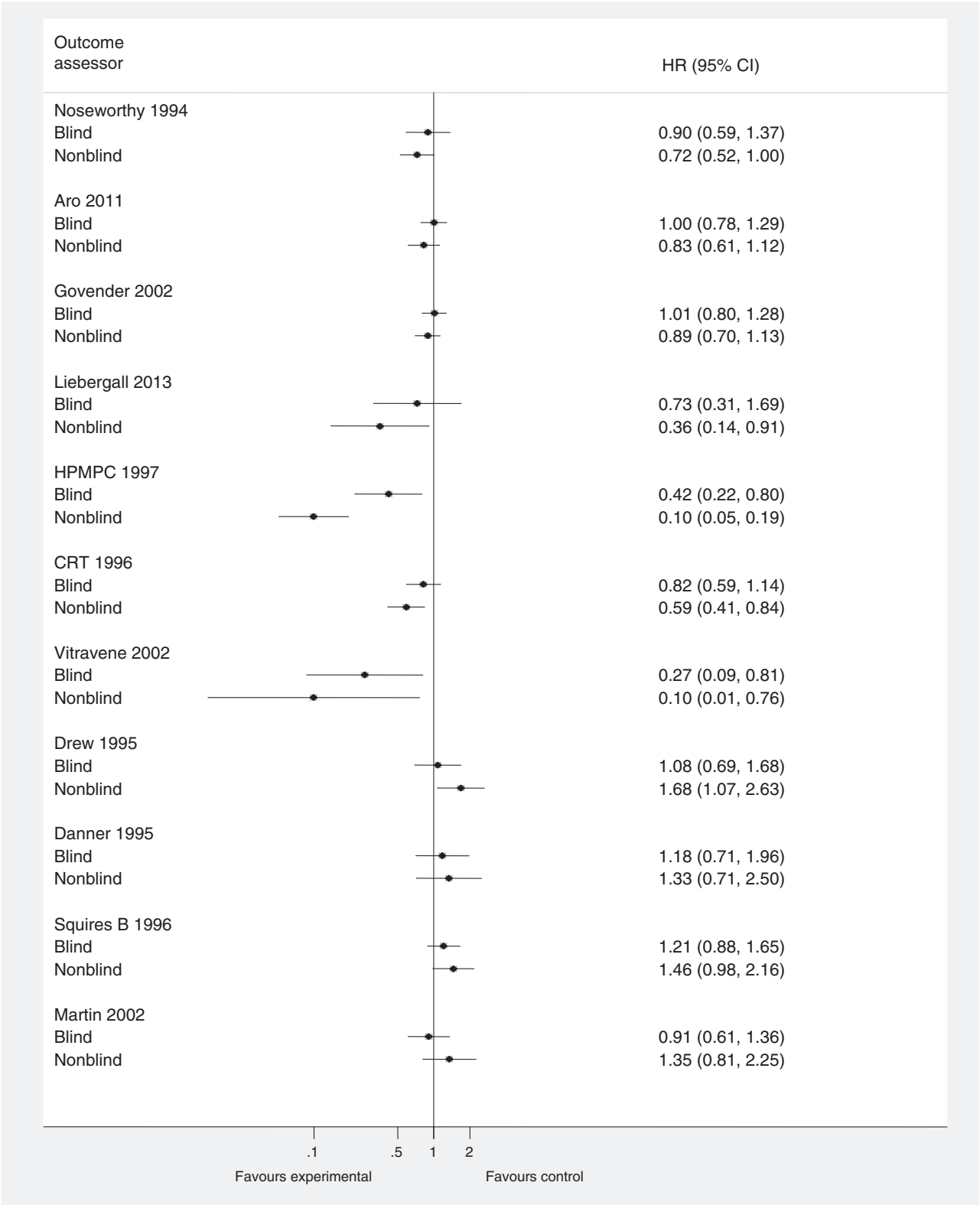

Figure 1. Estimated treatment effect in randomized trials with time-to-event outcomes according to type of outcome assessor: blind or nonblind. HR: hazard ratio.

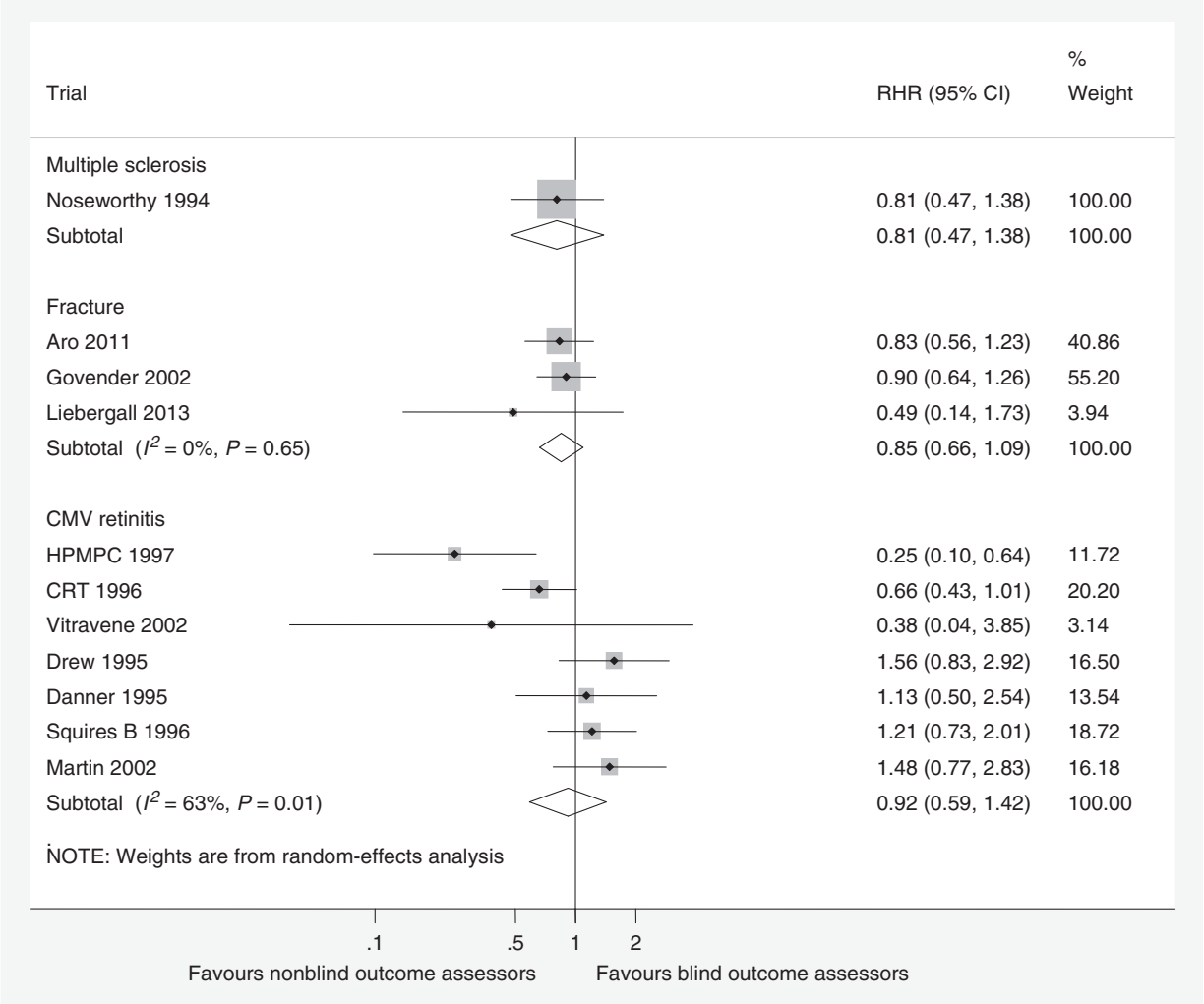

Figure 2. The impact of nonblind outcome assessors on the estimated treatment effect in randomized clinical trials with time-to-event outcomes—stratified by clinical condition. RHR: ratio of hazard ratios = $HR_{nonblind}/HR_{blind}$.

Qualitative summary of trials with incomplete or inconsistent data

Six trials (872 patients) were included in our review but not in our meta-analyses because of incomplete or inconsistent data. Two trials provided no data.[32,33]

Three trials provided some indication of observer bias in the expected direction.[28–30] Still in 2003[28] compared collagen sponge vs standard care in 82 burn patients. Kaplan–Meier curves were incompatible with tables, but disregarding this inconsistency, the ratio of hazard ratios was 0.73, indicating an exaggeration of the hazard ratio by 27%.

Swiontkowski 2006[29] compared human bone morphogenetic protein-2 vs standard care in 60 patients with tibial fractures. The trial was a smaller twin study to Govender 2002[18] which had a ratio of hazard ratios of 0.90, indicating an exaggeration of the hazard ratio by 10%.

Novo TFF[30] compared an electrical field device vs standard care in 237 patients with recurrent glioblastoma multiforme. If blinded assessment based on magnetic resoning imaging (MRI) only and nonblinded assessment based on both MRI and clinical examination were compared, the ratio of medians ratio was 1.18, which is approximately equivalent to an exaggeration of the hazard ratio by 18% (Appendix 3, available as Supplementary data at *IJE* online).

One trial indicated possible observer bias in the reversed direction.[31] Dumville 2009[31] compared larvae vs standard treatment for leg ulcers in 267 patients. Assuming time to censoring was identical for the blinded and nonblinded assessments, the ratio of hazard ratios was 1.10, indicating an underestimation of the hazard ratio by 10%.

Subgroup and sensitivity analyses

If the direction of bias was inverted in the four atypical cytomegalovirus retinitis trials, the ratio of hazard ratios was 0.75 (0.64 to 0.89) (Table 3). No trials were free from

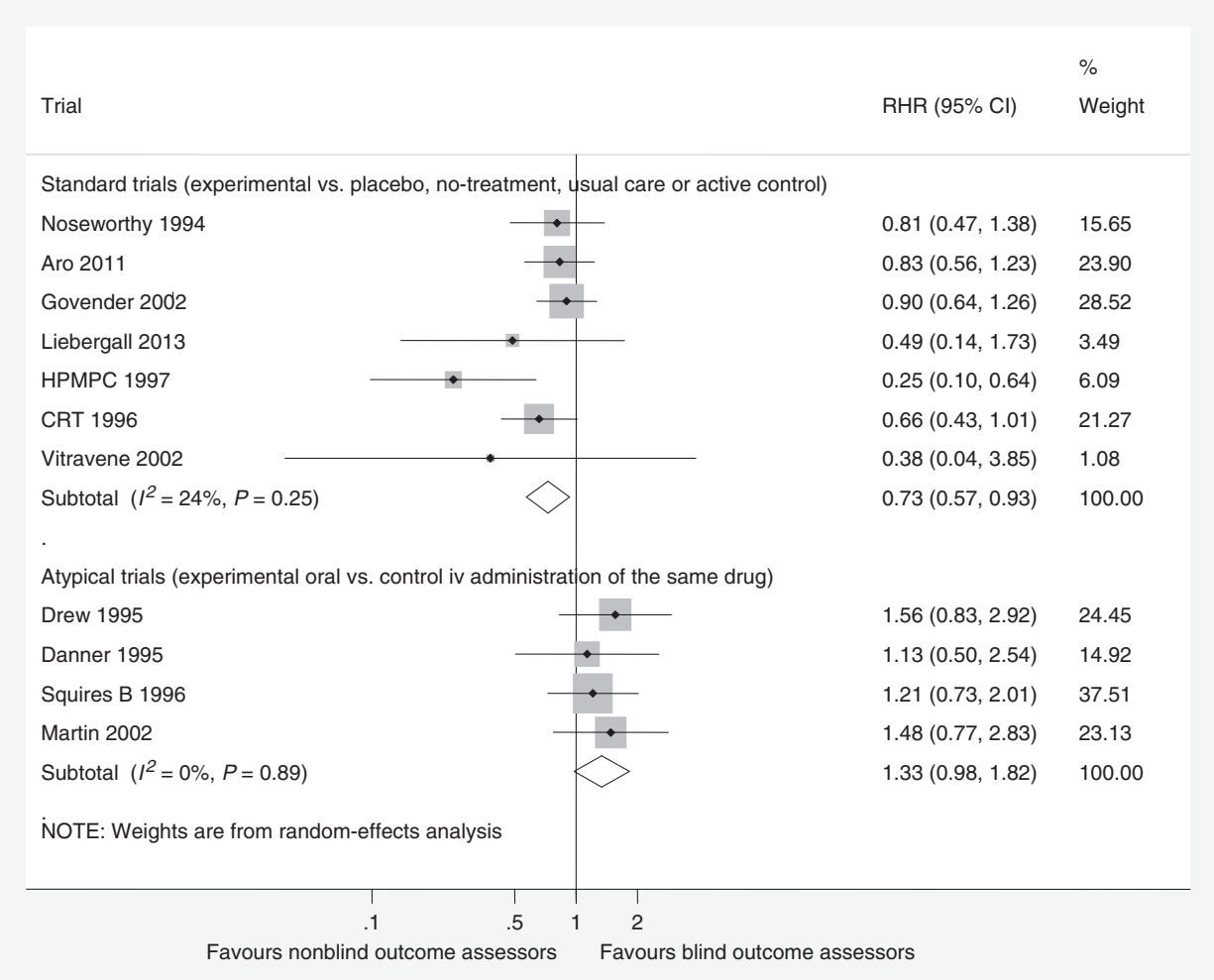

Figure 3. The impact of nonblind outcome assessors on the estimated treatment effect in randomized clinical trials with time-to-event outcomes—stratified by type of trial. RHR: ratio of hazard ratios = $HR_{nonblind}/HR_{blind}$.

all four possible measurement distortion factors, but they did not seem to be associated with the ratio of hazard ratio (Table 3).

Reasons for variation in observer bias

Informal inspection of the score values of the three outcome characteristics revealed no specific pattern (Appendix 4, available as Supplementary data at *IJE* online). Reversed direction of observer bias was a likely cause of observer bias variation among the cytomegalovirus retinitis trials.

Discussion

Failure to blind outcome assessors in 11 randomized time-to-event trials exaggerated hazard ratios by an average of 12% (-12% to 31%) but unconditional pooling of the trials was not meaningful due to qualitative heterogeneity. In seven standard trials involving cytomegalovirus retinitis, tibia fracture or multiple sclerosis, nonblinded outcome assessors favoured the experimental intervention, exaggerating the hazard ratio by an average of 27% (7% to 43%). In four atypical cytomegalovirus retinitis trials the direction of bias reverted to favour the control treatment by a comparable average degree of bias.

The strength of our study is the low risk of confounding in the direct comparison of the blinded and nonblinded assessors of the same outcome in the same trial involving the same patients. Three trials provided individual patient data with follow-up assessments after event occurrence, securing high data quality and enabling incorporation of the dependence between blinded and nonblinded assessments in our estimate of the ratio of hazard ratios.

A possible study limitation is heterogeneity. The difference in results between the two groups of cytomegalovirus retinitis trials was striking, with ratio of hazard ratios 0.46 (0.22 to 0.94) vs 1.33 (0.98 to 1.82), i.e. of comparable

Table 3. Main and secondary analyses

Comparisons	N	I^2 (*P*-value)	[a]RHR (95% CI)
Main analyses			
All trials	11	44% (0.06)	0.88 (0.69 to 1.12)
Standard trials (experimental vs placebo/no-treatment/usual care/active control)	7	24% (0.25)	0.73 (0.57 to 0.93)
Atypical [b]CMV trials(experimental oral vs control [c]iv administration of the same drug)	4	0% (0.89)	1.33 (0.98 to 1.82)
Supportive analyses			
Reversed direction of bias in atypicalCMV trials	11	0% (0.59)	0.75 (0.64 to 0.89)
All trials have equal weight	11	[d]NA	0.77 (NA)
Trials with individual patient data	3	0% (0.93)	0.86 (0.68 to 1.08)
[e]Dependence between blind and nonblind assessments accounted for			
paired (blind, nonblind) patient level data: trials with individual patient data	3	0% (0.81)	0.86 (0.75 to 0.98)
paired (blind, nonblind) patient level data/correction factor: standard trials	7	72% (0.01)	0.66 (0.51 to 0.85)
paired (blind, nonblind) correction factor: atypical CMV trials	4	0% (0.65)	1.34 (1.11 to 1.62)
paired (blind, nonblind) patient level data/correction factor: all trials	11	79% (0.00)	0.85 (0.67 to 1.08)
Trial characteristics (atypical CMV trials excluded)			
Nonblind assessment: multiple observer consensus	0	NA	NA
Nonblind assessment: single observer	7	24% (0.25)	0.73 (0.57 to 0.93)
Publication status: observer bias a main objective	1	NA	0.80 (0.47 to 1.38)
Publication status: observer bias not a main objective	6	36% (0.17)	0.70 (0.52 to 0.94)
Funding: non-industry or unclear	1	NA	0.80 (0.47 to 1.38)
Funding: industry	6	36% (0.17)	0.70 (0.52 to 0.94)
[f]Risk of measurement distortion (atypical CMV trials excluded)			
Timing of blind and nonblind assessment: same/similar	7	24% (0.25)	0.73 (0.57 to 0.95)
Timing of blind and nonblind assessment: not same/similar	0	NA	NA
Assessors: same type (e.g. neurologist vs neurologist)	2	0% (0.12)	0.75 (0.46 to 1.22)
Assessors: not same type (e.g. radiologist vs surgeon)	5	45% (0.12)	0.70 (0.50 to 0.97)
Procedure: same type (e.g. radiographs vs radiographs)	2	0% (0.56)	0.72 (0.51 to 1.00)
Procedure: not same type (e.g. photo vs live observation)	5	46% (0.12)	0.67 (0.45 to 1.02)
Blinding procedures: probably effective	4	0% (0.73)	0.84 (0.66 to 1.08)
Blinding procedures: possibly not effective	3	57% (0.10)	0.58 (0.35 to 0.98)

[a]Ratio of hazard ratios = hazard $\text{ratio}_{\text{nonblind}}$/hazard $\text{ratio}_{\text{blind}}$ (confidence interval); [b]cytomegalovirus; [c]intravenous; [d]not assessable or no data; [e]the standard error in the three trials with data on the dependence between blind and nonblind assessments (paired patient-level data) was reduced by a median of 40% when dependence was incorporated and this reduction was used as a correction factor in the seven trials without such data; [f]exploration of factors that potentially could distort the assessment of the difference between blind and nonblind trial result.

size but in opposite directions. This seems to be a rare example of a meta-analysis with a qualitative difference in the direction of an effect between two subgroups.[34] The meta-analysis of all trials had an I^2 of 44%, which does not normally disqualify meta-analysis, but the qualitative difference in the direction of bias among the cytomegalovirus trials and their corresponding I^2 of 63% precludes a meaningful pooling. Furthermore, the quantification of heterogeneity was underpowered as we had no access to a correlation factor appropriate for all trials. We previously noted[13] the likely reversed direction of observer bias in one cytomegalovirus trial;[26] however, we stress that the formal distinction between standard and atypical trials was not predefined for the present study.

The cytomegalovirus retinitis trials were from the 1990s when studies involving HIV patients were conducted in a highly politicized environment. The HIV epidemic, the initial lack of effective treatment and the subsequent testing of possibly effective drugs provoked a heated debate between gay activist organizations, religious zealots, physicians and researchers, drug producers and drug regulation agencies.[35] We find it credible that the clinicians in these trials preferred an established intravenous treatment, and expected that the experimental oral administration of the same drug, though more practical, might not be as beneficial. Thus, the heterogeneity can be explained by a simple design factor common to the atypical trials and coherent with likely bias mechanisms. The cytomegalovirus retinitis trials provide an illustrative case of how the clinical and wider public contexts of otherwise comparable scientific investigations shape the predispositions of its investigators and thereby the direction and degree of observer bias.

Searching for trials with both blinded and nonblinded assessors is challenging, and unidentified trials may exist. Still, publication bias is normally driven by the treatment effect,[36] so missed trials may, on average, have limited yet

unpredictable impact on our analyses of two types of assessment.

The number of trials in our review is moderate, and the number of clinical conditions low, so extrapolation to trials in general may be problematic. Still, the risk of observer bias probably has less to do with the specific clinical conditions than with trial settings. We included no trials with objective outcomes, for example time-to-death, so our results apply to trials with subjective outcomes. The included cytomegalovirus trials may reflect a worst-case situation, and pooling them with other trials is a matter of discussion. Future updates of our review may clarify whether the degree of observer bias in the cytomegalovirus trials (either favouring the experimental or the control intervention) is uncharacteristically large. Furthermore, extrapolation of our results to trials with subjective time-to-event outcomes in general assumes that trials with both blinded and nonblinded outcome assessors are comparable to trials with only nonblinded assessors.

Previous studies of observer bias in randomized clinical trials have found that nonblinded assessors, on average, exaggerated odds ratios by 36%[12] and standardized mean differences by 67%.[13] In combination, the three studies included 51 trials and 24 clinical conditions, and provide strong evidence of a marked risk of observer bias in randomized trials with nonblinded outcome assessor of subjective outcomes. Savović *et al*. pooled results from seven meta-epidemiological studies and found a 23% exaggeration of odds ratios in trials with subjective outcomes that were not double-blinded.[10] Their important study does not distinguish performance bias from observer bias, but their result is nonetheless compatible with ours.

In time-to-event trials the categorization of an event is often not based on a distinct dichotomization, as for example between life and death, but is juxtaposed upon a continuously developing biological process, for example the healing of a bone. The degree of subjectivity in assessing such events may thus be considerable, and will generate variation between different observations but not necessarily bias. Observer predisposition in the context of a subjective outcome is required to generate bias, which may thus vary unpredictably from observer to observer and from trial to trial. It is difficult to predict the direction or the size of any bias in an individual trial. We would thus not recommend that our pooled average is used as a form of correction factor. When ascertaining the possible bias in a trial with nonblinded assessors we would recommend consideration of the range of possible observer bias, and not only our pooled average, as well as the subjectivity of the outcome involved and possible cues for observer predispositions.

At odds with the suggestion of the US Food and Drug Administration,[37] the Pharmaceutical Research and Manufacturers of America's working group recommended against the general use of blindeded independent central review of cancer trials assessing progression-free survival, and suggested instead a sample based audit.[3] The working group included employees from six major pharmaceutical companies with an interest in improving the opportunity for getting drugs approved for marketing by regulatory authorities. The group's recommendation was in part based on a meta-analysis of 27 trials comparing local investigators' combined clinical/radiological assessments of progression-free survival with the central and blinded radiological assessments. The analysis reported an exaggeration of hazard ratios derived from local investigators by an average of only 3% (-2% to 8%).[38] However, this result should not be crudely interpreted as reflecting the impact of blinding, because in many included trials local investigators were blinded (and not only the assessors conducting the central review), and because the outcomes compared were not identical (clinical event locally and radiographical event centrally).

It is almost always possible to implement blinded outcome assessment in randomized trials, and creative solutions have been worked out in challenging cases.[39,40] Good inter-observer agreement does not seem to prevent observer bias,[13] so formal training of nonblinded observers[41] cannot safely replace blinding. Nor is a sample-based audit a good alternative,[3] because sample-based analyses tend to be underpowered. In trials with nonblinded clinicians and central blinded outcome assessment, informative censoring may be a concern, especially if blinded assessments stop once a patient is categorized as having an event by the nonblinded clinician. It is therefore important to continuee blinded follow-up assessments beyond that time point and, if possible, also to have the blinded assessments conducted in real time, impacting directly on the clinical decision process.[42]

When deciding whether to implement blinded outcome assessment in a trial, the main dilemma will often be that an intangible reduction of the risk of bias comes with a concrete increase in cost and trial complexity. Empirical methodological studies provide an estimate of the degree, range and direction of observer bias, and feed importantly into the decision whether to blinded or not, as well as informing the interpretation of results from conducted trials.

In conclusion, lack of blinded outcome assessors in randomized trials with subjective time-to-event outcomes causes a high risk of observer bias. Nonblinded outcome assessors in randomized clinical trials with subjective time-to-event outcomes tend to favour the experimental intervention and exaggerate hazard ratios by an average of 27% (7% to 43%); but in special situations, nonblinded outcome assessors favour the control intervention,

inducing a comparable degree of observer bias in the reversed direction.

Supplementary Data

Supplementary data are available at *IJE* online.

Funding

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

Acknowledgements

We thank the following researchers for sharing with us their unpublished outcome data: Peggy Vandervoort and George C Ebers, Alexandre Valentin-Opran, Nicky Cullum (individual patient data); and Jacob Lalezari (detailed outcome data). We also thank Peter C Gøtzshe for valuable comments to a previous version of the manuscript.

A.H. conceived the idea and design, organized the study and wrote the first draft of the manuscript. A.S.S.T and A.H. developed the search strategy. A.S.S.T., F.E., B.T., S.B., J.R. and A.H. carried out the nonmasked data collection. I.B., P.R., S.B., B.T. and A.H. carried out the masked data collection. A.H. did the statistical analysis on Stata, J.H. did the statistical analyses on SAS. All authors discussed the result and commented on the manuscript. A.H. had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Conflict of interest: None declared.

References

1. Haahr M, Hróbjartsson A. Who is blind in randomized clinical trials? An analysis of 200 trials and a survey of authors. *Clin Trials* 2006;3:360–65.
2. 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;90:190–96.
3. Amit O, Mannino F, Stone AM *et al*. Blinded independent central review of progression in cancer clinical trials: results from a meta-analysis. *Eur J Cancer* 2011;47:1772–78.
4. Noseworthy JH, Ebers GC, Vandervoort MK, Farquhar RE, Yetisir E, Roberts R. The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. *Neurology* 1994;44:16–20.
5. Higgins JPT, Green S (eds). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0. Cochrane Collaboration, 2011. www.cochrane-handbook.org (20 December 2013, date last accessed).
6. Burkhardt JE, Ennulat D, Pandher K *et al*. Topic of histopathology blinding in nonclinical safety biomarker qualification studies. *Toxicol Pathol* 2010;38:666–67.
7. Dodd DC. Blind slide reading or the uninformed versus the informed pathologist. *Comments Toxicol* 1988;2:88–91.
8. Stone AM, Bushnell W, Denne J *et al*. PhRMA working group. Research outcomes and recommendations for the assessment of progression in cancer clinical trials from a PhRMA working group. *Eur J Cancer* 2011;47:1763–71.
9. Pildal J, Hróbjartsson A, Jørgensen KJ, Hilden J, Altman DG, Gøtzsche PC. Impact of allocation concealment on conclusions drawn from meta-analyses of randomized trials. *Int J Epidemiol* 2007;36:847–57.
10. Savović 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.
11. Devereaux PJ, Manns BJ, Ghali WA *et al*. Physician interpretations and textbook definitions of blinding terminology in randomized controlled trials. *JAMA* 2001;285:2000–03.
12. Hróbjartsson A, Thomsen AS, Emanuelsson F *et al*. Observer bias in randomized clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. *BMJ* 2012;344:e1119.
13. Hróbjartsson A, Thomsen AS, Emanuelsson F *et al*. Observer bias in randomized clinical trials with measurement scale outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. *CMAJ* 2013;185:E201–11.
14. Tierney JF, Stewart LA, Ghersi D, Burdett S, Sydes MR. Practical methods for incorporating summary time-to-event data into meta-analysis. *Trials* 2007;8:16.
15. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7:177–88.
16. Noseworthy JH, Vandervoort MK, Penman M *et al*. Cyclophosphamide and plasma exchange in multiple sclerosis. *Lancet* 1991;337:1540–41.
17. Aro HT, Govender S, Patel AD *et al*. Recombinant human bone morphogeneticprotein-2: a randomized trial in open tibial fractures treated with reamed nail fixation. *J Bone Joint Surg Am* 2011;93:801–08.
18. Govender S, Csimma C, Genant HK *et al*. Recombinant human bone morphogenetic protein-2 for treatment of open tibial fractures: a prospective, controlled, randomized study of four hundred and fifty patients. *J Bone Joint Surg Am* 2002;84-A:2123–34.
19. Liebergall M, Schroeder J, Mosheiff R *et al*. Stem cell-based therapy for prevention of delayed fracture union: a randomized and prospective preliminary study. *Mol Ther* 2013;21:1631–38.
20. Parenteral cidofovir for cytomegalovirus retinitis in patients with AIDS: the HPMPC peripheral cytomegalovirus retinitis trial. A randomized, controlled trial. Studies of Ocular Complications of AIDS Research Group in Collaboration with the AIDS Clinical Trials Group. *Ann Intern Med* 1997;126:264–74.
21. Combination foscarnet and ganciclovir therapy vs monotherapy for the treatment of relapsed cytomegalovirus retinitis in patients with AIDS. The Cytomegalovirus Retreatment Trial. The Studies of Ocular Complications of AIDS Research Group in Collaboration with the AIDS Clinical Trials Group. *Arch Ophthalmol* 1996;114:23–33.
22. Vitravene Study Group. A randomized controlled clinical trial of intravitreous fomivirsen for treatment of newly diagnosed peripheral cytomegalovirus retinitis in patients with AIDS. *Am J Ophthalmol* 2002;133:467–74.

23. Drew WL, Ives D, Lalezari JP *et al*. Oral ganciclovir as maintenance treatment for cytomegalovirus retinitis in patients with AIDS. Syntex Cooperative Oral Ganciclovir Study Group. *N Engl J Med*. 1995;333:615–20.
24. The Oral Ganciclovir European and Australian Cooperative Study Group. Intravenous vs. oral ganciclovir: European/Australian comparative study of efficacy and safety in the prevention of cytomegalovirus retinitis recurrence in patients with AIDS. *AIDS* 1995;9:471–77.
25. Squires KE. Oral ganciclovir for cytomegalovirus retinitis in patients with AIDS: results of two randomized studies. *AIDS* 1996;10(Suppl 4):S13–18.
26. Martin DF, Sierra-Madero J, Walmsley S *et al*. A controlled trial of valganciclovir as induction therapy for cytomegalovirus retinitis. *N Engl J Med* 2002;346:1119–26.
27. Lalezari JP, Friedberg DN, Bissett J *et al*. Roche Cooperative Oral Ganciclovir Study Group. High dose oral ganciclovir treatment for cytomegalovirus retinitis. *J Clin Virol* 2002;24:67–77.
28. Still J, Glat P, Silverstein P, Griswold J, Mozingo D. The use of a collagen sponge/living cell composite material to treat donor sites in burn patients. *Burns* 2003;29:837–41.
29. Swiontkowski MF, Aro HT, Donell S *et al*. Recombinant human bone morphogenetic protein-2 in open tibial fractures. A subgroup analysis of data combined from two prospective randomized studies. *J Bone Joint Surg Am* 2006;88:1258–65.
30. http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/MedicalDevices/MedicalDevicesAdvisoryCommittee/NeurologicalDevicesPanel/UCM247246.doc (19 June 2012, date last accessed).
31. Dumville JC, Worthy G, Bland JM *et al*. Larval therapy for leg ulcers (VenUS II): randomized controlled trial. *BMJ* 2009;338:b773.
32. Musch DC, Martin DF, Gordon JF, Davis MD, Kuppermann BD. Treatment of cytomegalovirus retinitis with a sustained-release ganciclovir implant. The Ganciclovir Implant Study Group. *N Engl J Med* 1997;337:83–90.
33. Kabat-Zinn J, Wheeler E, Light T *et al*. Influence of a mindfulness meditation-based stress reduction intervention on rates of skin clearing in patients with moderate to severe psoriasis undergoing phototherapy (UVB) and photochemotherapy (PUVA). *Psychosom Med* 1998;60:625–32.
34. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA* 1991;266:93–98.
35. Singer M. The politics of AIDS. Introduction. *Soc Sci Med* 1994;38:1321–24.
36. Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ* 1997;315:640–45.
37. www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM071590.pdf (28 June 2012, date last accessed).
38. Dodd LE, Korn EL, Freidlin B *et al*. Blinded independent central review of progression-free survival in phase III clinical trials: important design element or unnecessary expense? *J Clin Oncol* 2008;26:3791–96.
39. Boutron I, Guittet L, Estellat C, Moher D, Hróbjartsson A, Ravaud P. Reporting methods of blinding in randomized controlled trials assessing non-pharmacological treatments. A systematic review. *PLoS Med* 2007;4:e61.
40. Karanicolas PJ, Bhandari M, Walter SD, Heels-Ansdell D, Guyatt GH. 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;62:214–23.e1.
41. Brorson S, Bagger J, Sylvest A, Hrobjartsson A. Improved interobserver variation after training of doctors in the Neer system. A randomized trial. *J Bone Joint Surg Br* 2002;84:950–54.
42. Meignan M, Itti E, Bardet S *et al*. Development and application of a real-time on-line blinded independent central review of interim PET scans to determine treatment allocation in lymphoma trials. *J Clin Oncol* 2009;27:2739–41.